VOLUME 103 PART B NUMBER 10 JULY 1956

The Proceedings

OF

THE INSTITUTION OF ELECTRICAL ENGINEERS

FOUNDED 1871: INCORPORATED BY ROYAL CHARTER 1921

PART B

RADIO AND ELECTRONIC ENGINEERING
(INCLUDING COMMUNICATION ENGINEERING)

SAVOY PLACE . LONDON W.C. 2

Price Seven Shillings and Sixpence

The Institution of Electrical Engineers

FOUNDED 1871
INCORPORATED BY ROYAL CHARTER 1921

PATRON: HER MAJESTY THE QUEEN

COUNCIL 1955-56

President

SIR GEORGE H. NELSON, Bart.

Past-Presidents

SIR JAMES SWINBURNE, Bart., F.R.S.
W. H. ECCLES, D.Sc., F.R.S.
THE RT. HON. THE EARL OF MOUNT EDGCUMBE, T.D.
J. M. DONALDSON, M.C.
PROFESSOR E. W. MARCHANT, D.Sc.
P. V. HUNTER, C.B.E.
H. T. YOUNG.
SIR GEORGE LEE, O.B.E., M.C.
SIR ARTHUR P. M. FLEMING, C.B.E., D.Eng., LL.D.
J. R. BEARD, C.B.E., M.Sc.
SIR NOEL ASHBRIDGE, B.Sc.(Eng.).
COLONEL SIR A. STANLEY ANGWIN, K.B.E., D.S.O., M.C., T.D., D.Sc. (Eng.).
SIR HARRY RAILING, D.Eng.
P. DUNSHEATH, C.B.E., M.A., D.Sc. (Eng.).
SIR VINCENT Z. DE FERRANTI, M.C.
T. G. N. HALDANE, M.A.
PROFESSOR E. B. MOULLIN, M.A., Sc.D.
SIR ARCHIBALD J. GILL, B.Sc.(Eng.).
SIR JOHN HACKING.
COLONEL B. H. LEESON, C.B.E., T.D.
SIR HAROLD BISHOP, C.B.E., B.Sc.(Eng.).
J. ECCLES, C.B.E., B.Sc.

Vice-Presidents

T. E. GOLDUP, C.B.E.
S. E. GOODALL, M.Sc.(Eng.).
WILLIS JACKSON, D.Sc., D.Phil., Dr.Sc.Tech., F.R.S.
SIR HAMISH D. MACLAREN, K.B.E., C.B., D.F.C., LL.D., B.Sc.
SIR W. GORDON RADLEY, K.C.B., C.B.E., Ph.D.(Eng.).

Honorary Treasurer

THE RT. HON. THE VISCOUNT FALMOUTH

Ordinary Members of Council

PROFESSOR H. E. M. BARLOW, Ph.D., B.Sc.(Eng.).
J. BENNETT.
C. M. COCK.
A. R. COOPER, M.Eng.
A. T. CRAWFORD, B.Sc.
B. DONKIN, B.A.
PROFESSOR J. GREIG, M.Sc., Ph.D.
F. J. LANE, O.B.E., M.Sc.
G. S. C. LUCAS, O.B.E.
D. MCDONALD, B.Sc.
C. T. MELLING, C.B.E., M.Sc.Tech.
H. H. MULLENS, B.Sc.
W. F. PARKER.
R. L. SMITH-ROSE, C.B.E., D.Sc., Ph.D.
G. L. WATES, J.P.
G. O. WATSON.
D. B. WELBOURN, M.A.
J. H. WESTCOTT, B.Sc.(Eng.), Ph.D.
E. L. E. WHEATCROFT, M.A.
R. T. B. WYNN, C.B.E., M.A.

Chairmen and Past-Chairmen of Sections

Measurement and Control:
W. BAMFORD, B.Sc.
*M. WHITEHEAD.

Radio and Telecommunication:
H. STANESBY.
*C. W. OATLEY, O.B.E., M.A., M.Sc.

Supply:
L. DRUCQUER.
*J. D. PEATTIE, C.B.E., B.Sc.

Utilization:
D. B. HOGG, M.B.E., T.D.
*J. I. BERNARD, B.Sc.Tech.

Chairmen and Past-Chairmen of Local Centres

East Midland Centre:
F. R. C. ROBERTS.
*J. H. MITCHELL, B.Sc., Ph.D.

North Midland Centre:
F. BARRELL.
*G. CATON.

North-Western Centre:
G. V. SADLER.
*PROFESSOR E. BRADSHAW, M.B.E., M.Sc.Tech., Ph.D.

Scottish Centre:
E. WILKINSON, Ph.D., B.Eng.
*J. S. HASTIE, B.Sc.(Eng.).

Mersey and North Wales Centre:
PROFESSOR J. M. MEEK, D.Eng.
*P. R. DUNN, B.Sc.

North-Eastern Centre:
A. H. KENYON.
*G. W. B. MITCHELL, B.A.

Northern Ireland Centre:
MAJOR E. N. CUNLIFFE, B.Sc.Tech.
*MAJOR P. L. BARKER, B.Sc.

South Midland Centre:
H. S. DAVIDSON, T.D.
*A. R. BLANDFORD.

Southern Centre:
L. H. FULLER, B.Sc.(Eng.).
*E. A. LOGAN, M.Sc.

Western Centre:
T. G. DASH, J.P.
*A. N. IRENS.

* Past-Chairmen.

RADIO AND TELECOMMUNICATION SECTION COMMITTEE 1955-56

Chairman

H. STANESBY

Vice-Chairmen

R. C. G. WILLIAMS, Ph.D., B.Sc.(Eng.).
J. S. MCPETRIE, Ph.D., D.Sc.

Past-Chairmen

C. W. OATLEY, O.B.E., M.A., M.Sc.
J. A. SMALE, C.B.E., A.F.C., B.Sc.

Ordinary Members of Committee

PROF. H. E. M. BARLOW, Ph.D., B.Sc.(Eng.).
F. S. BARTON, C.B.E., M.A., B.Sc.
A. J. BIGGS, Ph.D., B.Sc.
E. V. D. GLAZIER, Ph.D.(Eng.), B.Sc.
G. G. MACFARLANE, Dr.Ing., B.Sc.
B. N. MACLARTY, O.B.E.
H. PAGE, M.Sc.
W. ROSS, M.A.
L. RUSHFORTH, M.B.E., B.Sc.
T. B. D. TERRONI, B.Sc.
A. M. THORNTON, B.Sc.
F. WILLIAMS, B.Sc.

And

The President (*ex officio*).
The Chairman of the Papers Committee.
PROF. H. E. M. BARLOW, Ph.D., B.Sc.(Eng.) (representing the Council).
E. H. COOKE-YARBOROUGH (Co-opted Member).
BRIG. E. J. H. MOPPETT (representing the Cambridge Radio and Telecommunication Group).
J. MOIR (representing the South Midland Radio and Telecommunication Group).
D. H. THOMAS, M.Sc.Tech., B.Sc.(Eng.) (representing the North-Eastern Radio and Measurements Group).

The following nominees of Government Departments:
Admiralty: CAPTAIN G. C. F. WHITAKER, R.N.
Air Ministry: AIR COMMODORE G. H. RANDLE, R.A.F., B.A.
Department of Scientific and Industrial Research: B. G. PRESSEY, M.Sc.(Eng.), Ph.D.
Ministry of Supply: BRIG. J. D. HAIGH, O.B.E., M.A.
Post Office: CAPTAIN C. F. BOOTH, O.B.E.
War Office: COL. E. I. E. MOZLEY, M.A.

Secretary

W. K. BRASHER, C.B.E., M.A., M.I.E.E.

Deputy Secretary

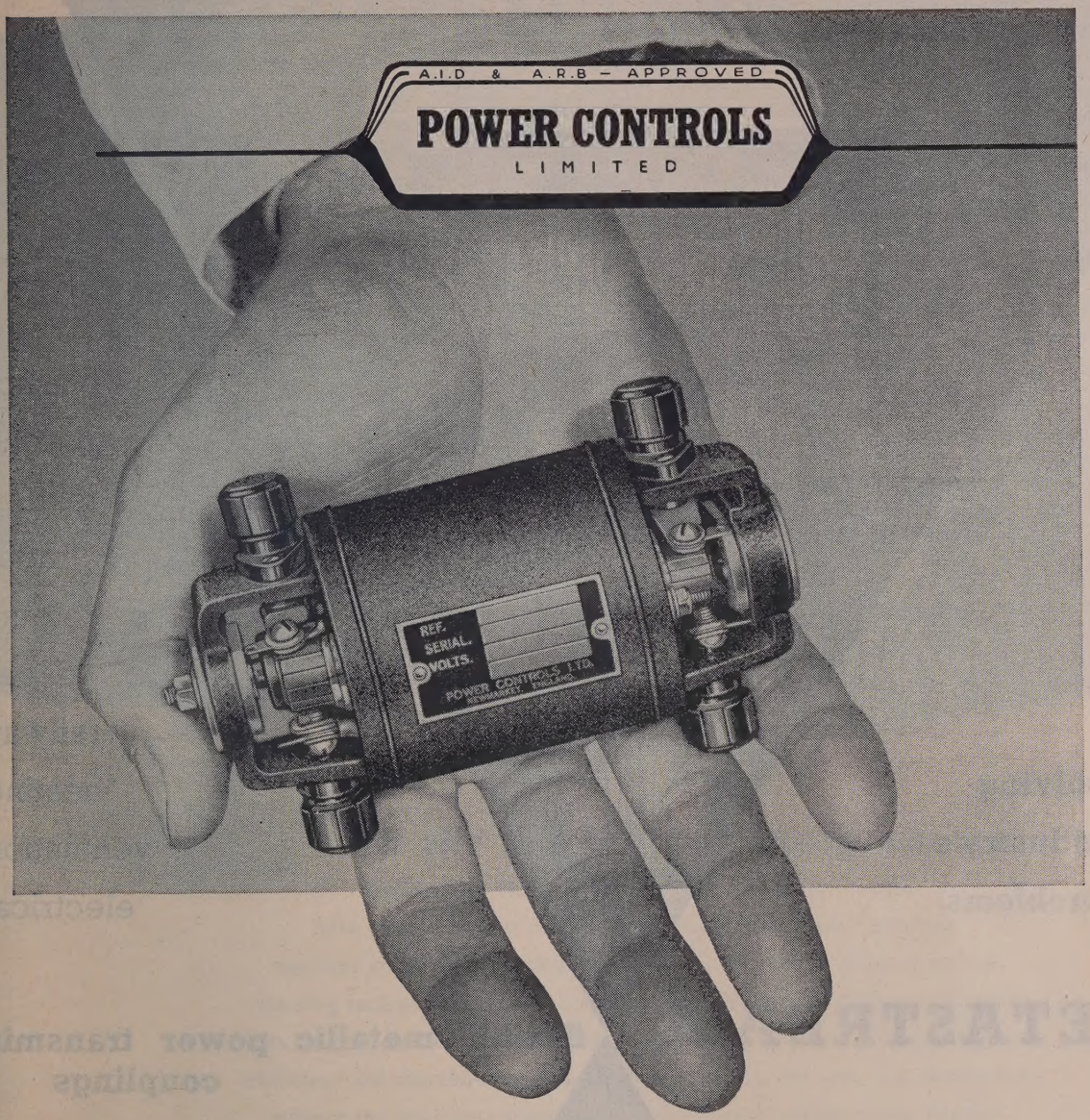
F. JERVIS SMITH, M.I.E.E.

Assistant Secretary

F. C. HARRIS.

Editor-in-Chief

G. E. WILLIAMS, B.Sc.(Eng.), M.I.E.E.



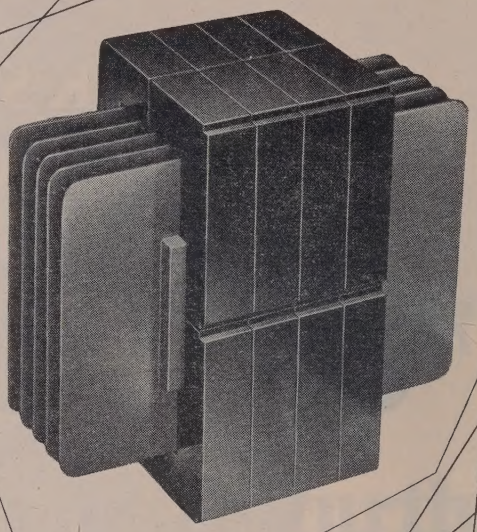
Rotary Transformers

Have you a transformer problem? If so, we can help you. We can undertake to develop and manufacture rotary transformers to your specification.

The illustration shows a typical transformer which we are manufacturing for a specific requirement. Made for 6, 12 or 24 volts D.C. input, it can supply a continuous D.C. output of 350 volts at 30 mA. or an intermittent output of 310 volts at 60 mA. The no-load current consumption is 2.2 amps. at 11.5 volts and the ripple voltage is less than 6 volts r.m.s. on 60 mA. load. The size is only 4-9/16″ long by 2-21/32″ across the brush terminals.

Power Controls Ltd., Exning Road, Newmarket, Suffolk

Telephone: Newmarket 3181. Telegrams: Powercon, Newmarket



H.F. POWER TRANSFORMERS

RATING . . . . . UP TO 2 kW

FREQUENCY RANGE . . . . 2 Kc/s to 2 Mc/s

H.F. power transformers of outstanding efficiency are the latest addition to the Mullard range of high quality components designed around Ferroxcube magnetic cores.

Utilising the unique characteristics of Ferroxcube to the full, Mullard H.F. transformers are smaller, lighter, and less costly than transformers using alternative core materials. These advantages are particularly marked in transformers required to handle powers of up to 2kW, between the frequency range 2kc/s to 2Mc/s.

Mullard transformers are already finding wide use in applications as diverse as ultrasonic H.F. power generators and aircraft power packs operating from an aircraft's normal A.C. supply. In the latter application, the low leakage field of Ferroxcube can eliminate the need for external screening, thereby reducing the size and weight of the transformer even further.

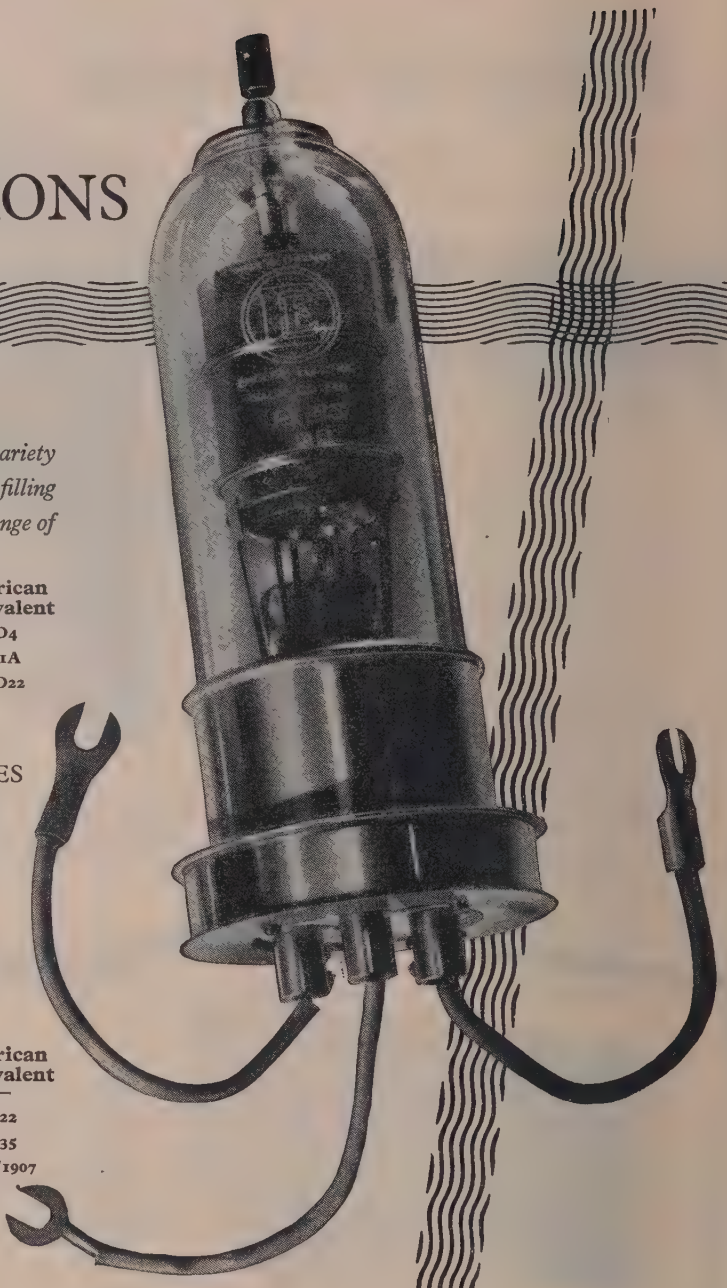
As with all Mullard high quality components, these H.F. power transformers are designed and built to engineers' individual specifications. Write now for details of the complete range of components available under this service.

Mullard

'Ticonal' permanent magnets
Magnadur ceramic magnets
Ferroxcube magnetic cores

MULLARD LTD · COMPONENTS DIVISION · CENTURY HOUSE · SHAFTESBURY AVENUE · WC2 MC1

THYRATRONS



XENON-FILLED TRIODES

These valves are suitable for a wide variety of control applications. The inert gas filling ensures stable operation over a wide range of ambient temperatures.

E.E.V. Type	C.V. Number	American Equivalent
AFX.212	1949	6D4
AFX.203	2868	C1A
CX.1113	2851	3D22

HYDROGEN-FILLED PULSE MODULATOR TRIODES

Designed to discharge pulse-forming networks in high power, high voltage pulse generators. Short deionization time and low time jitter provide for precise triggering at high repetition frequencies. Full technical data will be sent on request.

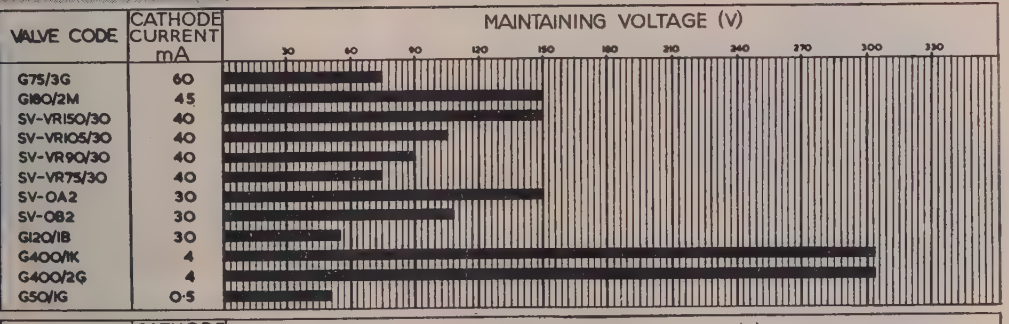
E.E.V. Type	C.V. Number	American Equivalent
FX.215	2203	—
FX.219	2520	5C22
FX.225	1787	4C35
FX.229	3521	5949/1907

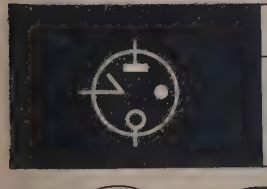
'ENGLISH ELECTRIC'

ENGLISH ELECTRIC VALVE CO. LTD. EE *Waterhouse Lane, Chelmsford*
Telephone: Chelmsford 3491

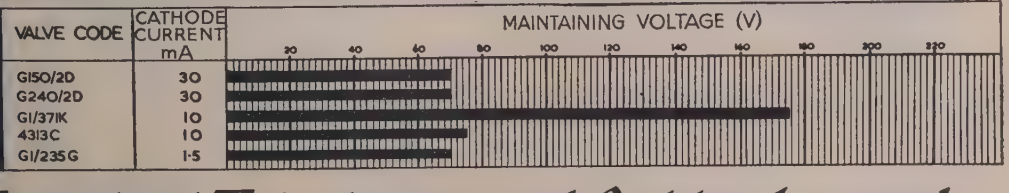
AP 300/39

VALVE CODE	CATHODE CURRENT mA
G75/3G	60
GI80/2M	45
SV-VR150/30	40
SV-VR105/30	40
SV-VR90/30	40
SV-VR75/30	40
SV-OA2	30
SV-OB2	30
GI20/1B	30
G400/1K	4
G400/2G	4
G50/1G	0·5





VALVE CODE	CATHODE CURRENT mA
GI50/2D	30
G240/2D	30
GI/371K	10
4313C	10
GI/235G	1·5



THE PROCEEDINGS OF THE INSTITUTION OF ELECTRICAL ENGINEERS

TEN-YEAR INDEX

1942—1951

A TEN-YEAR INDEX to the *Journal of The Institution of Electrical Engineers* for the years 1942–48 and the *Proceedings* 1949–51 (vols. 89–98) can be obtained on application to the Secretary.

The published price is £1 5s. 0d. (post free), but any member of The Institution may have a copy at the reduced price of £1 (post free).

This illustration shows an Arrow 30 amp Contactor actual size.

The smallest panel-mounting contactor on the market

50% saving in weight and size.

Complies with B.S.S. 775 for breaking capacity.

Coils and contacts changed in a matter of seconds.

Exceptionally low wattage consumption. C.S.A. approved.

Conforms with American N.E.M.A. specification.

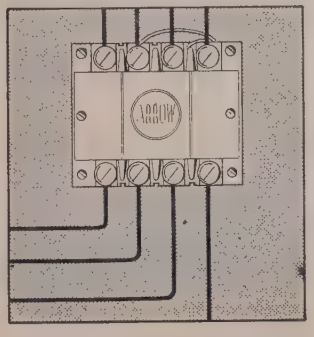
Comprehensive spares facilities in U.S.A. and Canada.

Three sizes — 30, 50 and 100 amps. at 550 volts A/C rating.

D/C ratings on request.

STRAIGHT-THROUGH WIRING

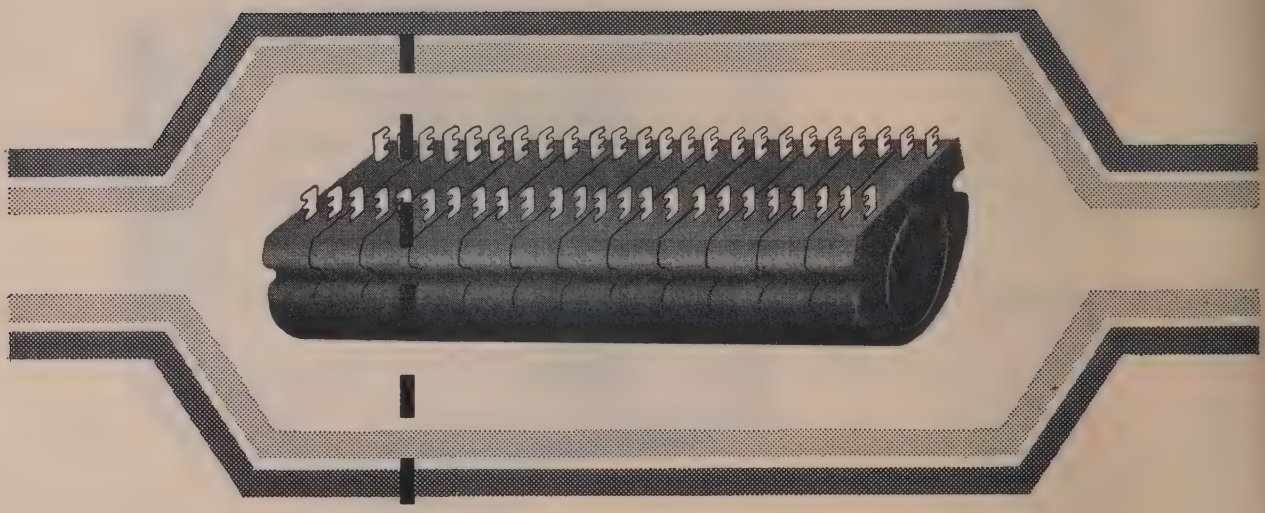
This is a completely new, built-in, advanced wiring design. Installation time is greatly reduced and circuit identification is easy and positive.



SEND FOR NEW CATALOGUE MS.9

ARROW ELECTRIC SWITCHES LTD · HANGER LANE · LONDON · W.5

Loading coils **inside** the cable splice

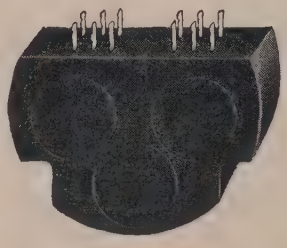


The advantages of the splice loading technique are particularly marked in the loading of small cables of up to 74 pairs. The coils can be included in a jointing sleeve or unit of only slightly larger diameter than would normally be used.

The loading coils in the Mullard L.160 Series are designed specifically for this technique. They are cast in resin, which provides complete protection from climatic conditions and allows a telephone administration to store them ready for building into loading units as and when required. Both single and triple assemblies are available for different sizes of cable.

Ferroxcube pot cores give these coils certain electrical advantages over conventional types, particularly in the loading of higher frequency circuits such as those encountered in programme and carrier applications.

You are invited to write for leaflets describing the Mullard L.160 Series coils and simple units for pole and splice loading.



TRIPLE COIL ASSEMBLY

Mullard

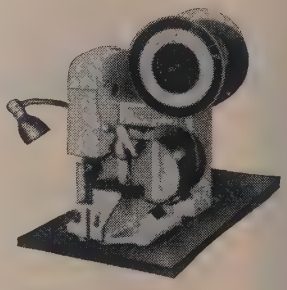
SPECIALISED ELECTRONIC EQUIPMENT

MULLARD LTD · EQUIPMENT DIVISION
CENTURY HOUSE · SHAFTESBURY AVENUE · W.C.2

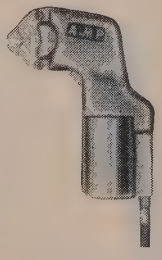
(MI453A)

Solderless terminals speed assembly eliminate rejects

The AMP precision method of wire termination reduces production costs and provides connections of the highest electrical and mechanical efficiency. The use of AMP precision crimping tools and automatic machines achieves exceptionally high rates of output, a uniformly high standard of quality and the elimination of human error. AMP terminations and connections are of particular value in electronics and aircraft installations. They withstand vibration, corrosion and provide high tensile strength, with low resistance, and no noise at R.F.



Automatic wire terminator operates at up to 4,000 an hour.



Pneumatic hand tool eliminates operator fatigue.

Certi-crimp hand tool. Positive closure ensured.

Brochure I.E.E. sent or demonstration at your own works on request.

AMP terminals are made for every type and size of wire.

AIRCRAFT-MARINE PRODUCTS (GT. BRITAIN) LTD.

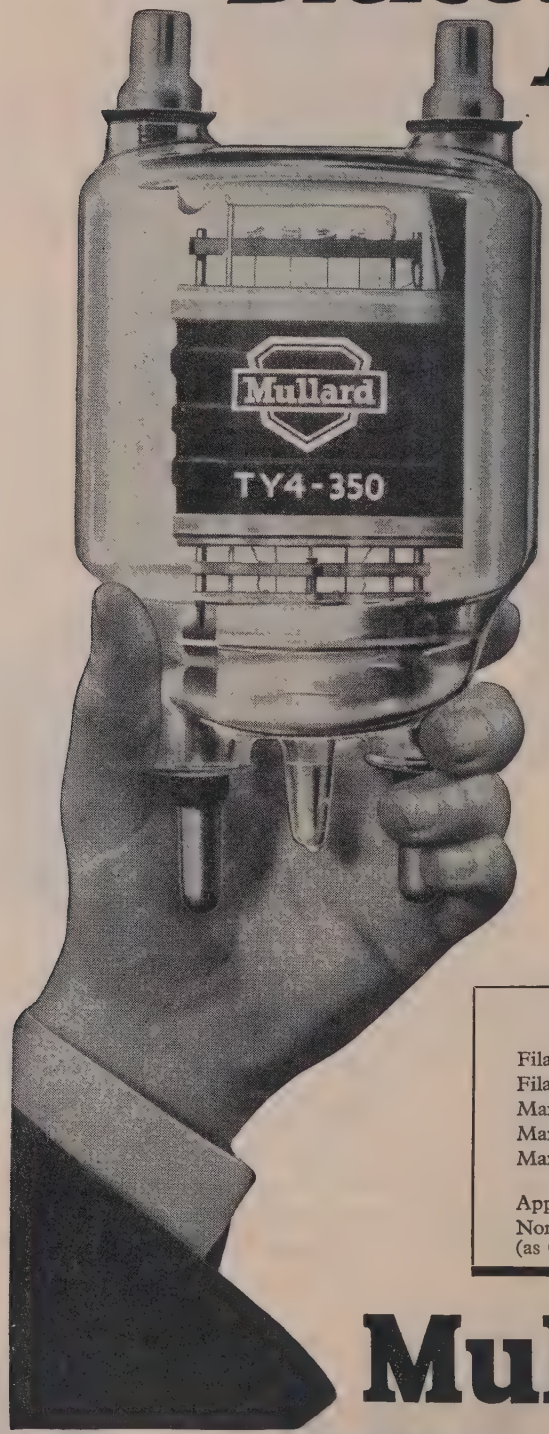
London Sales Office: 60 KINGLY STREET, LONDON, W.1. Tel: REGent 2517/8

Works: SCOTTISH INDUSTRIAL ESTATES, PORT GLASGOW, SCOTLAND

Ahead of the present — — abreast of the future

AP323-13

for Induction and Dielectric Heater Maintenance



TY4-350 (833A) POWER TRIODE

The Mullard power triode, TY4-350 (CV635) is a direct equivalent of the American 833A, and can be used with confidence as a plug-in replacement.

PRINCIPAL CHARACTERISTICS

	Radiation cooled	*Forced-air cooled*
Filament Voltage	10V	10V
Filament Current	10A	10A
Max. Anode Voltage	3000V	4000V
Max. Anode Dissipation	300W	400W
Max. Operating Frequency at reduced ratings	75Mc/s	75Mc/s
Approx. Power into matched load	800W	1200W
Nominal Power Output (as Class C oscillator at 20 Mc/s)	1000W	1400W

Mullard

MULLARD LTD., COMMUNICATIONS AND INDUSTRIAL VALVE DEPARTMENT, CENTURY HOUSE, SHAFTESBURY AVENUE, LONDON, W.C.2. MVT 188

The Institution is not, as a body, responsible for the opinions expressed by individual authors or speakers.
An example of the preferred form of bibliographical references will be found beneath the list of contents.

THE PROCEEDINGS OF THE INSTITUTION OF ELECTRICAL ENGINEERS

EDITED UNDER THE SUPERINTENDENCE OF W. K. BRASHER, C.B.E., M.A., M.I.E.E., SECRETARY

VOL. 103. PART B. NO. 10. JULY 1956

621.317.334.089.6:621.3.011.3

Paper No. 2080 M
July 1956

THE CALIBRATION OF INDUCTANCE STANDARDS AT RADIO FREQUENCIES

By L. HARTSHORN, D.Sc., Member, and J. J. DENTON, B.Sc.

(*The paper was first received 2nd December, 1955, and in revised form 16th February, 1956.*)

SUMMARY

The precision with which standards of inductance can be measured and used at radio frequencies is limited, not only by the ordinary uncertainties of experimental technique, but also by the fact that the standards are necessarily circuit elements with terminations, while inductance is primarily a characteristic of complete circuits. The procedure followed at the National Physical Laboratory when the highest possible precision is required is described in both its experimental and theoretical aspects. The precise relation between the familiar equivalent network for a coil and the basic definitions is indicated, and practical details are given for the resonance method of measurement that has been adopted as standard practice for all work of this kind. It is successfully operated at all frequencies from the audio range, where it overlaps the bridge methods employed as standard practice in the lower ranges of frequency, to values little short of those of self-resonance of the coils. In the overlapping region the bridge and resonance methods agree to about 1 part in 10^4 and with suitable elaboration the resonance method gives about the same accuracy for coils of any value between 1 H and 10 μH at any frequency within their working ranges, the limit for the smaller coils being about 4 Mc/s. With coils of the order of 1 μH measured values may show a standard deviation as small as 0·000 1 μH, but from the general considerations outlined in the paper limits closer than 0·000 5 μH will seldom be significant in any work with lumped circuits.

(1) INTRODUCTION

The paper describes the practice adopted at the National Physical Laboratory for the calibration of laboratory standards of inductance for use at radio frequencies. It will be shown that an accuracy of about 1 part in 10^4 is now obtained for standards covering a considerable range of values, but since the accuracy with which such standards can be measured and used is determined as much by limitations associated with the definition of inductance as by errors in the experimental observations, it will be necessary to discuss the definition of inductance in some detail before proceeding to a consideration of the experimental technique.

(2) THE DEFINITION OF INDUCTANCE

Inductance is a property that is rigorously applicable only to a complete electric circuit. The basic concept is that of the mutual inductance, M, of two closed filamentary circuits, i.e. circuits formed by conductors of negligibly small cross-section, and the defining relations are

$$\Phi_2 = Mi_1 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

$$e_2 = M\frac{di_1}{dt} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

where Φ_2 is the magnetic flux through circuit 2 produced by the current i_1 in circuit 1, and e_2 is the e.m.f. generated in circuit 2 by variations of i_1. The limitation to closed circuits is to be seen in the fact that Φ can have no definite value except for a closed contour; it is the surface integral of the magnetic induction vector over any surface bounded by the contour. Strictly speaking, the same limitation holds for the induced e.m.f., e: it is the line integral of the electric-force vector taken round a complete circuit.

It follows from eqns. (1) and (2) that, if currents i_1 and i_2 circulate in the filamentary circuits 1 and 2, the two circuits possess a mutual potential energy, W, given by

$$W = Mi_1i_2 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (3)$$

and also from Ampère's law that for two filamentary circuits in empty space

$$M = k\oint_1\oint_2 \frac{dl_1 dl_2}{r} \quad . \quad . \quad . \quad . \quad . \quad (4)$$

where dl_1 and dl_2 denote elements of the two circuits, r is the distance between them and k is a constant, not of necessity dimensionless, depending on the units employed.

The calibration of all inductance standards at the N.P.L. depends ultimately on eqn. (4) evaluated for the Campbell primary standard of mutual inductance. This standard conforms very closely to the postulated conditions; both the primary and secondary coils are practically closed, i.e. the leads in which they terminate are so closely spaced that the gaps between them, δl_1 and δl_2, if closed by wire links of these lengths, would contribute to eqn. (4) an amount that is negligible compared with M; secondly, both coils are wound of thin wire, the distance between them, r, being everywhere very large compared with the radius of the wire, so that the approximation obtained by treating them as filamentary circuits located in the axis of the wire can be

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.
The paper is an official communication from the National Physical Laboratory.

shown by application of eqn. (4) to be adequate for all practical purposes. Thus the Campbell standard with its value calculated in this way by eqn. (4) serves as a concrete realization of the unit of inductance of the highest precision so far obtainable.

It is well known, however, that standards of mutual inductance have comparatively little direct application to work at radio frequencies; for practical reasons standards of self-inductance are almost universally employed, and the special features on which the accuracy of the Campbell standard depends are of necessity abandoned. The working standard for use at radio frequencies consists of a single coil, wound with a conductor, usually of considerable cross-section in order to keep down the resistance, and not nearly closed but ending in two terminals, with a spacing large enough to ensure that the capacitance between them is constant and not unduly large. It is now necessary to consider the significance of the term 'inductance' applied to such a coil.

The basic conception of inductance will obviously not be applicable until the circuit is completed in some way or other, and common sense suggests that the best general way is to connect the terminals together by means of a straight link of a cross-section roughly similar to that of the winding conductor. In this way the contributions of the link to eqn. (4) and to the contour to which Φ of eqn. (1) is applicable are kept small and at least as precisely defined as the contributions of the winding itself. It is natural to take as a first approximation to the closed coil a filamentary circuit located along the axis of the wire, but inductance has no meaning in relation to a *single** filamentary current, and the basic conception becomes applicable only if the coil is considered as a bundle of such currents distributed over the cross-section of the wire in a way that corresponds approximately to the actual distribution of current density, which can usually be deduced with the required approximation from considerations of resistivity, skin-effect, etc. Every pair of such filaments, say a and b, has a mutual inductance M_{ab} determined by eqns. (1)–(4), and therefore the coil as a whole has a potential energy W which is the sum of the values for every pair of filaments

$$W = \tfrac{1}{2}\sum_{a=1}^{a=n}\sum_{b=1}^{b=n} M_{ab} i_a i_b \quad . \quad . \quad . \quad . \quad (5)$$

The $\tfrac{1}{2}$ appears because, by allowing both a and b to take all possible values in the summations, the term for each pair of filaments is included twice. Since in practice the only measurable current is the total current $i = \Sigma i_a = \Sigma i_b$, we must, following O'Rahilly,[1] use the relation between W and i for our working concept of the inductance of a single circuit.

If the cross-section, A, of the wire is uniform and if $\delta A'$ represents an element of A enclosing the filament i', and the current-density over $\delta A'$ is J', then we must have $i' = J'\delta A'$, and for any other filament i'' and its corresponding element of area $\delta A''$ and current density J'', $i'' = J''\delta A''$. Thus the above summation can be replaced by an integral over the uniform cross-section:

$$W = \tfrac{1}{2}\int\int MJ'dA'J''dA'' \quad . \quad . \quad . \quad . \quad (6)$$

where M denotes the mutual inductance of the two filaments enclosed by $\delta A'$ and $\delta A''$ respectively. If J_m is the mean current density over the whole cross-section A, then $J_m A = i$, and eqn. (6) can therefore be written in the form

$$W = \tfrac{1}{2}Li^2 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (7)$$

* The assumption commonly made, that eqns. (11)–(13) may serve to determine the inductance of even a single filament, merely leads in this case to a value which is infinite and therefore not measurable, the flux density becoming infinite at points on a current-carrying conductor when its radius is reduced to zero.

where $$L = \int\int M\frac{J'}{J_m}\frac{dA'}{A}\frac{J''}{J_m}\frac{dA''}{A} \quad . \quad . \quad . \quad . \quad (8$$

It will be clear from this equation that the quantity L, which w define as the self-inductance of a closed circuit of finite cross section, is of the same dimensions as M: it is merely the sum o a distribution of M's; both M and L are therefore legitimatel termed inductances and measured in terms of the same unit, an both are calculable from linear dimensions for circuits of simpl geometrical form in empty space.

Eqns. (7) and (8) have comparatively little application i electrical measurements until they are supplemented by equation corresponding to (1) and (2). Consider the system of eqn. (6) The filamentary current $J''\delta A''$ induces a flux $MJ''\delta A'' = \delta\Phi$ i the other filament, and thus we may write

$$W = \tfrac{1}{2}\int\int d\Phi J'dA'$$

$$= \tfrac{1}{2}\int \Phi J'dA' = \tfrac{1}{2}\int \Phi\frac{J'dA'}{J_m A}i = \tfrac{1}{2}\Phi_m i \quad . \quad . \quad (9$$

where Φ represents the total flux through one filament due t the currents in all the rest, and Φ_m is the mean flux defined by

$$\Phi_m = \int \Phi\frac{J'}{J_m}\frac{dA'}{A} \quad . \quad . \quad . \quad . \quad . \quad (10$$

By eqns. (7) and (9)

$$\Phi_m = L_i \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (11$$

This equation corresponds to eqn. (1).

Any variation of i necessarily involves a corresponding varia tion in all the filamentary currents i', i'', and therefore in al the fluxes Φ, and thus the filamentary current i' is opposed by a e.m.f. $e' = d\Phi/dt$, induced by all the other filamentary currents The mean of these e.m.f.'s is given by

$$e_m = \int \frac{e'J'dA'}{J_m A} = \frac{d\Phi_m}{dt} = L\frac{di}{dt} \quad . \quad . \quad . \quad (12$$

the integral being taken over the cross-section as before. Th energy W of the whole circuit necessitates that the total curren i shall be associated with an opposing e.m.f. of this magnitude and the use of standards of self-inductance depends mainly o the application of this equation. However, e.m.f.'s are, i general, measurable only in terms of the p.d. required to balanc them, and since for a self-inductor any such p.d. is inseparabl from Ri, the p.d. which in accordance with Ohm's law is asso ciated with the resistance, R, of the conductor, the genera working equation for an inductive coil must be written

$$v = Ri + L\frac{di}{dt} \quad . \quad . \quad . \quad . \quad . \quad . \quad (13$$

The inductance L retains its full significance discussed abov only when this equation is limited to circuits that are sensibl complete; v is then the voltage that is in principle measurabl across any gap of negligible width made in such a circuit. Bu since the quantities v, R and i are applicable with no less precisio to parts of circuits as well as to complete circuits, it is possibl to evaluate L by means of eqn. (13) for any circuit elemen Eqn. (13) used in this way becomes a generalized definition o inductance applicable to circuit elements and consistent wit the basic definition for complete circuits. This definition i implicitly adopted in most practical measurements of inductanc

(3) MEASUREMENT BY DIFFERENCE

Since inductance is strictly determinate only for a complete circuit, the value for an open coil must usually be obtained from two measurements, one made on a complete circuit including the coil, and the other on the same circuit after the coil has been disconnected and the resulting gap closed with as little disturbance as possible. Eqn. (13) then indicates that the difference of the two observed values of L will give the inductance of the coil with as close an approximation as is obtainable. The inductance of any link inserted to close the circuit for the second measurement should obviously be allowed for in some way, and it is important also to remember that L in eqn. (13), unlike the ohmic resistance R, is not simply additive for circuit elements connected in series. If, for example, two circuit elements are sufficiently complete to have values which can be separately measured as L_1 and L_2, the series combination will have the value $L_1 + L_2$ only if they are located so that the flux, Φ_m, through each is entirely unaffected by the current in the other, i.e. when their mutual inductance, M_{12}, is zero. The values of L obtained by applying eqn. (13) to the composite circuit will in the general case be $L_1 + M_{12}$ and $L_2 + M_{12}$, with $L_1 + L_2 + 2M_{12}$ as the overall value.

Thus the inductance of a given coil is not a constant characteristic of that coil in all circumstances. If the value measured in isolation is L_i, say, the increment in value observed when the coil is connected in series with another circuit is $L_i + 2M$, where M is the mutual inductance between the coil and the rest of the circuit, and of this increment the portion allocated to the coil by eqn. (13) is $L_i + M$, the additional M being allocated to the rest of the circuit. The accuracy with which any such coil can be measured and used will obviously be limited by M, which is usually unknown and subject to variation when the coil is transferred from one measuring circuit to another.

In work on standards M is, of course, always made as small as possible; the coil is well separated from any part of the measuring circuit for which the flux-linkage Φ_m could be considerable, and the leads connecting the coil to this circuit are either of bifilar or coaxial form. For low-frequency work closely spaced and twisted flexible leads are used, and it is possible by simply reversing the lead connections to reverse the sign of M without appreciably changing its magnitude. The mean of the two readings so obtained gives L_i alone. In radio-frequency work, however, the capacitance in the circuit becomes a major consideration, and such reversals are impracticable, because of the changes of capacitance associated with them. The complications introduced by capacitance into the measurement of inductance must next be considered.

(4) SELF-CAPACITANCE

The use of eqn. (13) to determine the inductance of a coil with respect to currents of a radio frequency is complicated by the fact that the observed current is not simply the current flowing within the conductor, but includes also capacitance current flowing between neighbouring turns of the coil. Consider any two turns of the winding that are adjacent in space though not necessarily successive turns in the winding; they may be in adjacent layers and separated by several other turns in the order of winding. The capacitance between two such turns will be considerable, especially if they are closely spaced and the conductor is of large diameter. As a first approximation the capacitance, although actually distributed around the turns, may be represented by a lumped capacitance C linking the mid-points of the two turns AB and EF (Fig. 1). Consider first the effect of any one such capacitor on the values of R and L determined by

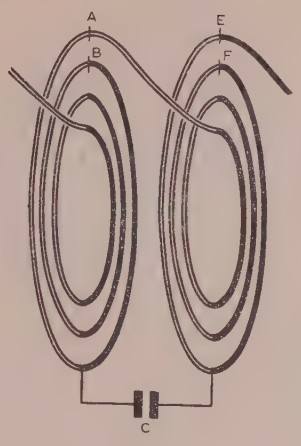


Fig. 1

applying eqn. (13) to the coil and its parts. The capacitor spans a portion of the coil for which the equation may be written

$$V_1 = (R_1 + jL_1\omega)I_1$$

and therefore the capacitor itself must carry a current $I_c = j\omega CV_1$. Thus the current in the remainder of the coil is $I_2 = I_1 + I_c$ and the equation for this part of the coil becomes

$$V_2 = (R_2 + jL_2\omega)I_2$$

In accordance with Section 3 the values of L_1 and L_2 include any mutual inductance between the two parts. If now the coil is considered as a single circuit-element its voltage is $V = V_1 + V_2$ and the current $I = I_2$; its characteristic equation becomes

$$\frac{V}{I} = \frac{V_1 + V_2}{I_2} = R_2 + jL_2\omega + \frac{R_1 + jL_1\omega}{1 - L_1C\omega^2 + j\omega CR_1} \quad (14)$$

Eqn. (14) shows immediately that the conditions that must be satisfied in order that eqn. (13) may adequately represent the properties of the coil are that the frequency must be sufficiently low to ensure that both $L_1C\omega^2 (= \gamma_1$, say) and $R_1C\omega (= \beta_1$, say) are negligible compared with unity. At radio frequencies this condition is not satisfied and therefore it becomes necessary to consider how the deviations from the basic equation shall be dealt with in making measurements of inductance.

It is, of course, possible from observations made at any one angular frequency ω to derive values of L and R that satisfy eqn. (13) applied to the overall voltage and current. Such values would not be true inductances and resistances, but would correspond to the imaginary and real terms of eqn. (14);

$$L = L_2 + \frac{L_1(1 - \gamma_1) - R_1\beta_1/\omega}{(1 - \gamma_1)^2 + \beta_1^2} \quad . \quad . \quad . \quad (15)$$

$$R = R_2 + \frac{R_1}{(1 - \gamma_1)^2 + \beta_1^2} \quad . \quad . \quad . \quad . \quad (16)$$

Two of the main features in the design of standard inductors for use at radio frequencies are that C is made as small as is practicable and R_1 is made as small as possible compared with $L_1\omega$. It follows that in all cases of practical interest γ_1 is less than unity but not necessarily small, $\delta_1 = R_1/L_1\omega$ is small compared with unity and $\beta_1 = \gamma_1\delta_1$ is still smaller, while β_1^2 and $R_1\beta_1/L_1\omega = \delta_1\beta_1$ are small quantities of the second order and negligible in

eqns. (15) and (16), which may therefore be used in the approximate forms

$$L \simeq L_2 + L_1/(1-\gamma_1)$$
$$R \simeq R_2 + R_1/(1-\gamma_1)^2$$

When all the capacitances between pairs of turns are taken into consideration these equations become

$$L \simeq \Sigma\frac{L_n}{1-\gamma_n} = \Sigma L_n + \Sigma\frac{L_n\gamma_n}{1-\gamma_n} = L_b + \Sigma\frac{L_n^2C_n\omega^2}{1-\gamma_n} \quad (17)$$

$$R \simeq \Sigma\frac{R_n}{(1-\gamma_n)^2} = R_b + \Sigma\frac{R_nL_nC_n\omega^2(2-\gamma_n)}{(1-\gamma_n)^2} \quad . \quad (18)$$

where L_b and R_b denote the inductance and resistance of the whole coil as defined by eqn. (13).

At radio frequencies inductance is mostly measured in terms of capacitance; the inductor is connected to a capacitor to form a resonant circuit and observations are made of corresponding values of the angular frequency, ω_r, at resonance and the total capacitance, C_x, external to the coil. The condition of resonance is that of zero reactance or $LC_x\omega_r^2 = 1$, which by eqn. (17) may be written

$$\frac{1}{\omega_r^2} = L_bC_x + C_x\omega_r^2\Sigma\frac{L_n^2C_n}{1-\gamma_n}$$

or

$$\frac{1}{\omega_r^2} = L_b\left[C_x + \Sigma\frac{L_n^2C_n}{L_bL(1-\gamma_n)}\right] \quad . \quad . \quad . \quad (19)$$

It will be shown later that the observed values of $1/\omega_r^2$ and C_x for a well-designed standard coil lie on a straight line with great precision. It follows from eqn. (19) that in such cases the slope of this line towards the C-axis is the basic inductance L_b, and the negative intercept on this axis is the quantity $\Sigma\frac{L_n^2C_n}{L_bL(1-\gamma_n)}$ which must be independent of frequency to the accuracy of the measurement. The nature of the term and the conditions in which it will be independent of frequency may be seen by putting it in the form

$$C_L \simeq \Sigma\frac{L_n^2(1-\bar{\gamma}_n)C_n}{L_b^2(1-\gamma_n)} \quad . \quad . \quad . \quad . \quad . \quad (20)$$

which is obtained by substituting $L = L_b/(1-\bar{\gamma}_n)$, a relation which is identical with eqn. (17) when $\bar{\gamma}_n$ denotes the mean value of γ_n given by

$$\bar{\gamma}_n = \frac{\Sigma\gamma_nL_n/(1-\gamma_n)}{\Sigma L_n/(1-\gamma_n)}$$

Each term in eqn. (20) is now clearly seen as a fraction of one of the capacitances between pairs of turns of the coil. At low frequencies, for which $\gamma_n = L_nC_n\omega^2$ is small compared with unity, the fraction is L_n^2/L_b^2 which is independent of frequency. At higher frequencies the γ terms will become appreciable, and since they increase as the square of the frequency, some variation of C_L with frequency may be expected when the frequency approaches that for which $\bar{\gamma}_n = 1$, the frequency at which the capacitance between turns is sufficient to establish resonance. However, the form of eqn. (20) shows that this variation of C_L with frequency must always be very slow, since γ_n and $\bar{\gamma}_n$ necessarily always vary together in a similar way, and if $\gamma_n \simeq \bar{\gamma}_n$ the low-frequency value of C_L will hold good at all frequencies up to that of self-resonance mentioned above. The condition required for a constant C_L is that each of the n elements of the coil associated with one of the dominant capacitances of the winding shall have about the same value of γ, i.e. C_n shall be distributed in such a way as to make L_nC_n constant throughout the winding. This condition is approximately satisfied in standard coils by regularity in the spacing and distribution of successive turns over the cross-section of the winding. Such coils are usually calibrated at frequencies up to about a quarter of that of self-resonance ($\bar{\gamma} = \frac{1}{16}$) and eqn. (19) is almost invariably found to be linear within the experimental error. Measurements at higher frequencies up to and including that of self-resonance are sometimes made, and it is usually found that no significant change in C_L with frequency can be detected even at these frequencies.

The quantity C_L, the self-capacitance of the coil, is clearly an important constant for a standard of inductance, and must be associated with the inductance L_b in order to determine the reactance of the standard at any frequency within its working range.

Three points of practical importance should be noted:

(*a*) The inductance, L_b, is independent of frequency only over ranges in which there is no appreciable change of current distribution as skin effect. There will be two such ranges: one at low frequencies, at which skin effect is negligible, and the other at high frequencies, at which the current is entirely concentrated in a thin surface layer. The value of L_b will in accordance with eqn. (8) be lower for the higher-frequency range.

(*b*) The self-capacitance, C_L, will not be independent of frequency unless the capacitance between turns is independent of frequency, which will be the case only when the wire insulation and coil former are of low-loss dielectrics. A change in earthing conditions may slightly alter the value of the total capacitance between turns and therefore affect C_L, but this must be distinguished from an effect of frequency.

(*c*) Eqns. (17) and (18) show that the relation between the overall voltage and current of the coil is the same as that for the circuit elements shown in Fig. 2. The effect of the coil on any circuit to which it is connected may therefore be completely represented by using Fig. 2 as the equivalent circuit of the coil. An equivalent network of this form has, of course, long been used on an empirical basis, but the relation between the postulated circuit elements and the fundamental laws of inductance, capacitance and resistance seems not to have been developed in detail, and it is essential that this should be done if such coils are to serve as standards.

Fig. 2

(5) EXPERIMENTAL METHOD

(5.1) General

As mentioned above, the method employed for the measurement of inductance at radio frequencies consists in connecting the coil to be measured to a standard variable capacitor, observing the total capacitance in parallel with the coil, and the corresponding frequency of resonance for a series of settings of the capacitor, and so determining the coefficients in the linear equation between $1/\omega_r^2$ and C_x. These coefficients give what may reasonably be termed the true or basic self-inductance of the coil and its self-capacitance. It should be noticed that the procedure amounts essentially to the determination of the reactance of the coil, at a series of frequencies covering its working range, by reference to a working standard of capacitance and a frequency standard. The standard of capacitance is used as a transfer standard, since it has been calibrated at low frequencies by reference to the primary standard of mutual inductance and the calibration is then assumed to be valid also at the radio frequencies—an assumption

hat is justified provided that the dielectric of the capacitor is ensibly free from power loss and the plates of the capacitor, i.e. he conductors conveying the current from the terminals to the lielectric do not themselves possess appreciable inductance. The mount of such inductance in the capacitor used can be measured by the same technique, i.e. observations of the frequencies of esonance of the circuit formed by short-circuiting the terminals of the capacitor. Details are given in a later paragraph after the main components of the apparatus have been considered.

The apparatus must include a generator capable of providing a series of frequencies of adequate stability, a standard capacitor of high precision provided with leads by means of which it can be connected to a standard coil to form a circuit of a precisely determinable resonant frequency, and a detector of resonance. The procedure adopted by the authors consists in setting the generator to induce into the measuring circuit, comprising coil, capacitor and detector of resonance, a constant e.m.f., and then varying the capacitance until the setting at which resonance is established is obtained. The capacitance C_x and the angular frequency ω_r are then noted, and the procedure is repeated for different frequencies until the required range has been covered. Special features required in these items in order to obtain the highest possible accuracy will now be discussed.

(5.2) The Generator

Very little power is required from the generator, but the frequency must be constant and known to a high order of accuracy. Both frequency and amplitude must be completely unaffected by the adjustments made to the measuring circuit in order to arrive at the resonance setting. Crystal control of the frequency clearly provides the best means of obtaining constancy of frequency; constancy of amplitude and a resulting constancy of e.m.f. induced in the measuring circuit can be simply obtained by using very loose inductive coupling between generator and measuring circuit, an electrostatic screen being placed around the inducing coil of the generator so as to eliminate any capacitive coupling which might change with manipulation of the measuring circuit.

A very convenient form of generator for work of this kind and one which is readily assembled from apparatus available in most standards laboratories was first introduced by Mr. W. Wilson some years ago, primarily for testing wavemeters. A double-beam cathode-ray oscillograph is used both to generate and monitor the oscillations, the various frequencies required being obtained as selected harmonics and sub-harmonics of a standard frequency with which the whole system is controlled. The arrangement is shown diagrammatically in Fig. 3. The signal

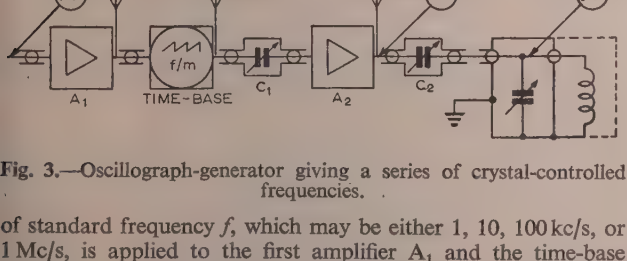


Fig. 3.—Oscillograph-generator giving a series of crystal-controlled frequencies.

of standard frequency f, which may be either 1, 10, 100 kc/s, or 1 Mc/s, is applied to the first amplifier A_1 and the time-base generator of the oscillograph, as shown, and the time-base is synchronized to any convenient submultiple, f/m, of the standard frequency. Thus the Y_1-plates of the oscillograph are energized at the frequency f and the X_1-plates at f/m. The time-base oscillations pass to the second amplifier, A_2, through a capacitor C_1 which serves as output control and is adjusted so as to overload A_2 and so produce even more harmonics in its output than are supplied from the time-base circuit. The gain of A_2 may also be varied to control the output, which then passes by way of another small capacitor C_2, to a circuit of high selectivity which can be tuned to any desired harmonic nf/m. This selector circuit consists of the inducing coil of the generator in parallel with a variable capacitor. The electrostatic screen to eliminate capacitive coupling consists of a wire cage surrounding the coil and connected to the screen of the capacitor. The stationary patterns seen on the oscillograph indicate the establishment of synchronism and the number of loops in the pattern indicates the value of the integer m. Thus, having first chosen a suitable value of f/m, a whole series of frequencies nf/m evenly spaced at intervals of f/m and known with the full accuracy of the frequency standard f is obtainable simply by adjusting the capacitor in the tuned circuit. The frequency interval f/m is, of course, chosen to suit the work in hand. For example, if the series of frequencies required is 1 000, 1 020, 1 040, 1 060, etc., kc/s, a frequency standard f of 100 kc/s is used, the time-base is synchronized at 20 kc/s or $f/5$, and the 50th, 51st, 52nd, etc., harmonics of this frequency are in turn selected. The key points, like 1 000 kc/s, 1 100 kc/s, can, of course, be checked by synchronizing the time-base at 100 kc/s. Frequencies from 800 c/s to 20 Mc/s, all having the accuracy of the crystal standard, are conveniently obtained in this way.

In order to screen the whole system, all the terminals on the oscillograph except the earth terminal were replaced by concentric screened plugs and the connections throughout were made with coaxial cable, the outer conductors forming part of the screen.

(5.3) Capacitors and Leads

The capacitance connected in parallel with the coil to be measured is provided by standard variable air capacitors of ordinary design fitted with rigid coaxial leads and terminations to which the coil and the detector can be conveniently connected. Rigidity of the leads is, of course, necessary to ensure that the capacitance of the leads themselves—which forms part of the total capacitance in parallel with the coil—shall remain constant. The coaxial form is used partly because the screening it affords is necessary to ensure that the capacitance has a precisely determinate value, and partly because the mutual inductance between the coil and the leads is negligibly small for such leads, and thus the uncertainty discussed in Section 3 is minimized. The capacitors themselves, if placed quite near the coil, might introduce appreciable mutual inductance or might change the magnetic flux and therefore the inductance of the coil by the mere proximity of so much metal. The coaxial leads are therefore made about 50 cm long, the reaction between coil and capacitor at this distance being negligible.

Two variable air capacitors are used, one for the higher frequencies having a maximum value of about 1 000 $\mu\mu$F and readable to 0·01 $\mu\mu$F at all points on its scale, and another for lower frequencies of maximum value 5 000 $\mu\mu$F readable to 0·2 $\mu\mu$F. When additional capacitance is required it is obtained by the use of a set of fixed mica capacitors having values between 5 000 and 60 000 $\mu\mu$F in steps of 5 000 $\mu\mu$F. These capacitors can be connected in parallel with the larger air capacitor either singly or in pairs, by connecting the capacitors between fixed terminals on a pair of parallel rigid copper bars, one attached to each of the terminals of the air capacitor. This arrangement ensures that the inductance of the copper connectors is definite and reproducible and the same for all the mica units within the limits of experiment, since the differences in the linear dimensions of the units are negligible. The screens of all the capacitors and leads included in the circuit at any one time are, of course,

connected together and to earth (unless the working frequency is so high that the earth capacitance of the whole screen has a lower impedance to earth than any earth conductor available).

(5.4) Detector

Various forms of detector can be used. High sensitivity is, of course, essential if full advantage is to be taken of the versatility of the generator, and it is also important that the detector shall add as little as possible to the total resistance, inductance and capacitance of the circuit. The most generally satisfactory detector used by the authors is a very sensitive voltage indicator consisting of a cathode-follower probe energizing a communication receiver, the probe being in series with an air-gap to form a shunt of capacitance about $0{\cdot}02\,\mu\mu$F and negligible conductance across the coil and capacitor. A signal-strength meter of adjustable sensitivity fitted to the receiver, and operated at a point which gives high sensitivity at readings near the resonance peak, provides a very convenient visual indicator, the maximum reading indicating the state of resonance.

For frequencies below 15 kc/s an audio-frequency amplifier followed by a rectifier microammeter takes the place of the radio receiver. A capacitor of about $2\,\mu\mu$F may be required in series with the input to the amplifier to obtain adequate sensitivity. A capacitance of this amount is usually of little consequence at audio frequencies, but in any case the capacitance of detectors of these two forms is easily measured and allowed for. However, in order to make the capacitance quite definite, the detector coupling device, probe, etc., must be located within the screen of the measuring capacitor or its leads. The detector probe is, in practice, mounted in a cylindrical screen which makes a T-joint with the coaxial lead between coil and capacitor, or else is plugged into a concentric termination on the capacitor screen.

It is to be noted that both of these detectors respond to the voltage across the coil and capacitors, and that it is therefore the setting for voltage or parallel resonance that is observed. This is slightly different from the current or series resonance referred to previously, but it is immaterial which is observed provided that the appropriate equation is used in calculating L_b. If ω_r is the angular frequency for voltage resonance with the procedure here described, the quantities in the equivalent circuit of Fig. 2 are given by

$$\frac{1}{\omega_r^2} = L_b\left(1 + \frac{1}{Q^2}\right)(C_x + C_L) = L_a(C_x + C_L) \quad . \quad (21)$$

where $Q^2 = L_b^2\omega_r^2/R_b^2$. For a standard inductor in its working range Q usually exceeds 100 and therefore the term $1/Q^2$ is negligible and $L_a = L_b$. When necessary, however, Q must be determined either from the sharpness of the resonance or by measuring R_b on a resistance bridge, and an appropriate correction is applied to L_a, the apparent value determined by the procedure previously outlined.

(5.5) Observational Procedure

The generator and measuring circuit are set up at a suitable distance apart, say 1 m between the inducing coil of the generator and the coil to be measured. The time-base oscillator is set to the required frequency-interval; the selector circuit is then set to give the first frequency required for the measurements, and the measuring circuit and the receiver are tuned to give maximum deflection with the generator output adjusted to a convenient level. It is, of course, necessary to check that the observed signal is from the measuring circuit only and that the measuring circuit is energized by the inducing coil only. When this is the case the observed resonance peak is symmetrical, and very precise settings can be made by taking the mean of the two capacitance readings that give the same signal on either side of the peak. The capacitance reading and the frequency having been noted, the process is repeated for the next frequency in the series, and so on until the whole range has been covered. The actual values of C_x at the resonance settings are then measured by disconnecting the coil and connecting the leads, capacitor and detector combination to a Schering capacitance bridge in an adjoining room by means of screened coaxial cables connected so that the capacitances of the cables are excluded from the measurements. The accuracy obtained by this procedure was noticeably better than could be obtained by taking the components to the bridge for calibration. The calibration is made at 1 kc/s for convenience, but the values can be proved to be independent of frequency at all frequencies up to the point at which the effects of internal inductance become appreciable. This will be considered later.

(5.6) Example

The series of corresponding values of $1/\omega_r^2$ and C_x is tabulated and the linear equation which best fits the observations is calculated by a method (such as Awbery's[2]) which gives values of L_a and C_L very simply, and provided that the coil has constant values of these quantities over the range of frequency covered by the observations, an uncertainty less than 1 in 10^4 for L_a and $0{\cdot}01\,\mu\mu$F for C_L is obtained. A region of constant inductance is usually most conveniently obtained by working at the highest attainable frequencies, i.e. with the smallest capacitor; the skin effect is then fully developed at all the frequencies used and the value of L_b obtained is the stationary limiting value for high frequencies. The value of C_L obtained in this way may be assumed to hold good at all frequencies (unless the coil is known to incorporate dielectric material of poor quality, in which case it must be rejected as a standard); the value of L_b for any other frequency can then be found from C_L and the observed values of ω_r and C_x for that frequency. The whole series of values of L_b obtained in this way provides the best indication of the accuracy of the values and of the variation of L_b with varying current distribution in the range of frequency, depending mainly on the diameter of the wire, at which the transition from uniform distribution to concentration in the surface layers occurs. Table 1 is a typical series of observations.

Table 1

TYPICAL OBSERVATIONS

Frequency	$1/\omega_r^2$	C_x	$C_x + C_L$	L_a
kc/s	$\times 10^{-12}$	$\mu\mu$F	$\mu\mu$F	μH
47·5	11·226 72	1 107·03	1 123·00	9 997·1
55	8·373 65	821·68	837·65	9 996·6
62·5	6·484 56	632·69	648·66	9 996·9
72·5	4·819 08	466·11	482·08	9 996·4
82·5	3·721 62	356·30	372·27	9 997·1
92·5	2·960 44	280·16	296·13	9 997·1
105	2·297 53	213·86	229·83	9 996·6
120	1·759 05	159·98	175·95	9 997·4
			Mean:	9 996·9
By Awbery's method: $L_a = 9\,996{\cdot}8\,\mu$H, $C_L = 15{\cdot}97\,\mu\mu$F			Standard deviation:	0·3

(6) CORRECTIONS FOR UNWANTED INDUCTANCE

The procedure described above gives the inductance of the whole measuring circuit, and this unavoidably includes a certain

amount of unwanted inductance in the leads and capacitor plates. The method of making due allowance for these inductances so as to obtain a value representative of the coil alone will now be considered.

The most obvious method of measuring the total residual inductance is to replace the coil by the shortest possible copper link of large cross-section and then to measure the inductance of this residual circuit by the same method. The measurements must, of course, be made at much higher frequencies because of the smallness of the inductance, but with fairly long leads the method is practicable, although it is usually more convenient to determine the frequency of resonance of circuits of this type by observation of the reaction of the circuit on an uncontrolled oscillator of variable frequency to which it is coupled. The accuracy of this method is quite adequate for the present purpose, for an approximate value of the residual inductance is all that is needed. If ω_0 is the observed angular frequency at resonance when C is the capacitance in the circuit, the residual inductance is given approximately by $l_r = 1/C\omega_0^2$, and the value of l_r can be obtained in this way for any setting of the capacitor. If the value of l_r proves to be independent of C, the method of difference of Section 3 is directly applicable; the value of l_r is deducted from the total circuit inductance first measured, the result giving the inductance of the coil less the inductance of the short copper link included in l_r; if necessary, the inductance of the link can be estimated by calculation from its dimensions and applied as a further correction.

Unfortunately this simple procedure fails, except as a first approximation, which is adequate only when measuring coils of large inductance, for two reasons: the measured inductance l_r is usually found to vary with C, and thus no single value applicable to the original series of observations is obtained; also l_r is measured at very high frequencies and can be used at much lower frequencies only if it is known that it represents the true inductance of real conductors in the system; the fact that it varies with C suggests that it is the equivalent value for a complex network of inductances and capacitances and may therefore vary considerably with frequency. These difficulties are overcome by finding, by a process of trial and error based on the known mechanical construction of the capacitors and leads, an equivalent circuit in which each inductance represents a real conductor of the structure and each capacitance a real dielectric path between conductors. Such inductances and capacitances may justifiably be taken as independent of frequency to the accuracy required, and if the network does, in fact, represent all the observations made on the residual circuit at the higher frequencies, it may be safely used to calculate the corresponding properties at the lower frequencies for any capacitor setting used. The equivalent circuits that have been obtained in this way for the capacitors and leads previously mentioned will serve to illustrate the procedure that is followed in such cases.

(6.1) The Equivalent Circuit for a Variable Air Capacitor

A circuit which has been found to represent adequately the variable air capacitors in most measurements is shown in Fig. 4(*a*); a slightly more complex circuit which may be needed when the coil to be measured is of very small inductance, so that a correspondingly higher accuracy is required in the corrections for capacitor inductance, is that shown in Fig. 4(*b*). In both cases the capacitance, $C - C_f$, associated with the moving plate is distinguished from that, C_f, associated with fixed conductors only. In Fig. 4(*a*) the whole of the inductance is treated as being associated with the moving plate and its counterpart in the fixed plate-system; in Fig. 4(*b*) the inductance of the terminal connection and fixed-plate system is treated separately. Actual

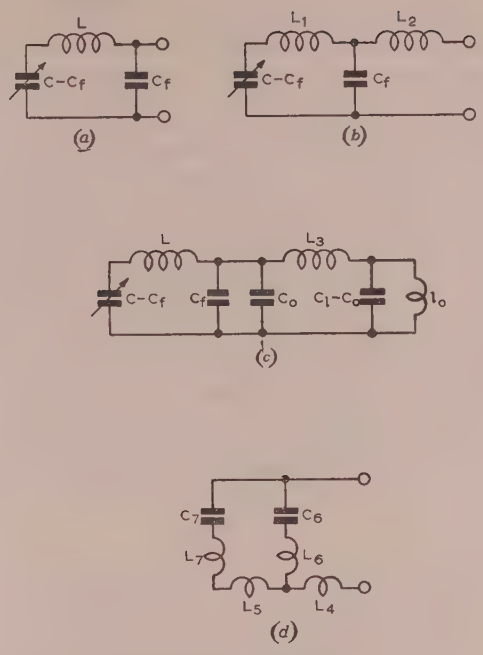


Fig. 4.—Equivalent networks for the residual circuit.
(*a*) and (*b*) Capacitor at setting C.
(*c*) Capacitor with leads and link.
(*d*) Mica capacitors.

Table 2

INDUCTANCE VALUES FOR A STANDARD CAPACITOR

Measured values		Equivalent circuit elements
C	l_r	
$\mu\mu$F	μH	
		Fig. 4(*a*)
5 000	0·065 72	$C_f = 29\,\mu\mu$F. $L = 0{\cdot}0669\,\mu$H (standard deviation, 0·000 6μH)
4 000	0·066 03	
2 000	0·066 23	*Fig.* 4(*b*)
1 000	0·065 10	$C_f = 60\,\mu\mu$F. $L_1(\mu\text{H}) = 0{\cdot}033\,86 - 0{\cdot}5 \times 10^{-6}C$
500	0·062 12	
300	0·058 10	$L_2 = 0{\cdot}035\,00\,\mu$H (standard deviation, 0·000 06 μH)

values found to represent one of the standard capacitors mentioned earlier are given in Table 2.

The standard deviation recorded is that for differences between the measured value of l_r and the corresponding value calculated for the network. These deviations show that the simple network shown in Fig. 4(*a*) is adequate for measurements on coils exceeding 10 μH, but the tenfold higher accuracy obtainable with the circuit shown in Fig. 4(*b*) becomes significant when measuring coils of, say, 1 μH. The slight change of the real inductance L_1 with C in Fig. 4(*b*) presumably arises from the change in the current distribution in the capacitor plates when they move.

(6.2) Equivalent Circuit for the Complete Residual Circuit

The equivalent circuits for the capacitors are found by observing the resonance of the capacitors alone, the only added element being a short-circuiting link, the inductance of which is allowed for by calculation; the equivalent circuit for the complete

residual circuit including the concentric leads and short-circuiting link is obtained by adding to the circuit shown in Fig. 4(*a*) elements based on direct measurements of the capacitance of the leads at, say, 1 kc/s, the calculated inductance, l_0, of the link and the observations of resonance of the complete residual circuit. The circuit obtained is shown in Fig. 4(*c*), in which C_l is the measured capacitance of the leads; C_0 is a portion of the leads capacitance which is found by trial to make the circuit accurately representative of all the observed resonance frequencies for a fixed value of L_3, C being the only variable parameter of the network. The circuit thus obtained, when used to evaluate the corrections to the main measurements of inductance, is often found to yield values that are consistent with one another within the limits set by random errors in the main observations, i.e. the results obtained for a given coil with different capacitors and at different frequencies are completely consistent with one another and with the basic equations, and this is the final justification for the whole procedure.

An example of the circuit shown in Fig. 4(*c*) evaluated for a coaxial lead and short-circuiting link used for this work is given in Table 3.

Table 3

RESIDUAL INDUCTANCE AND CAPACITANCE OF LEADS

Link: Copper strip, 1·9 cm × 0·16 cm × 4 cm long.
Leads: Concentric, copper with polystyrene spacers.
Length, 65 cm.
Inner, solid rod 0·47 cm diameter.
Outer, tube 1·91 cm outer diameter, and 1·59 cm inner diameter.
$l_0 = 0{\cdot}0150\,\mu$H. $C_l = 40{\cdot}4\,\mu\mu$F. $C_0 = 19\,\mu\mu$F.
$L_3 = 0{\cdot}1939\,\mu$H (standard deviation, $0{\cdot}0002\,\mu$H).

(6.3) Equivalent Circuit for the Mica Capacitors

Capacitances larger than 5000 $\mu\mu$F can be introduced into the measuring circuit by connecting fixed mica capacitors between fixed junction points on two rigid copper bars, which can be attached to the terminals of the variable capacitor. Allowance must, of course, be made for the inductance of the copper bars and the internal inductance of the mica capacitors. These quantities can be determined by a further application of the procedure described above, namely, observations of the resonance of the whole residual circuit including one or more of the mica capacitors. The form of the extension required to the circuits shown in Figs. 4(*a*), 4(*b*) or 4(*c*) is obvious. Thus, in the most complicated case, when two mica units are connected in circuit, the additional network is of the form shown in Fig. 4(*d*), where L_4 and L_5 represent inductances of parts of the copper bars and L_6 and L_7 represent inductances of the mica units. Obviously L_5 and L_7 can be lumped together as L_{57} and the three inductances L_4, L_6, L_{57}, which, when associated with the known capacitances C_6 and C_7 of the mica units, will account for the observed resonances of the composite residual circuit, can be found by trial as before. The values of C_6 and C_7 for the mica units are obtained by direct measurement at frequencies sufficiently low to ensure that the effect of inductance is negligible, usually 1 kc/s and 10 kc/s. The small diminution of capacitance with rise of frequency associated with dielectric loss must, of course, be allowed for when using the capacitors at higher frequencies; the necessary correction was estimated from the diminution of capacitance observed between 1 and 10 kc/s and the variation of tan δ with frequency. The consistency of the values of L_4, L_6, etc., for the same copper bars and different mica units, and of the values of L_b for the same coil obtained by using different capacitors, provides an adequate check of all the measurements.

(6.4) Procedure with Complex Residual Circuits

It will be appreciated that, when it is found to be necessary to use one of these complex circuits to account for residual inductance external to the coil, the procedure to be followed in calculating the inductance of the coil from the observations must be slightly modified; the linear equation between $1/\omega_r^2$ and C_x is no longer strictly true if C_x is the value given by the normal calibration of the capacitor, i.e. the actual capacitance of the dielectric path between its plates. It will, however, still be true if C_x is interpreted as the equivalent capacitance of the residual network measured at the coil terminals at the actual working frequency. This equivalent capacitance must therefore be calculated for each capacitor setting using the simplest of the networks that will serve the purpose in hand. The most complete of the circuits could, of course, always be used, but the arithmetic, although simple, becomes laborious when corrections for several residual inductances must be applied to every observation of capacitor setting.

(7) RESULTS

Values obtained by these methods for a range of standard inductors including the unit of each decade from 1 H to 10^{-6} H are given in Fig. 5. The scale, regarded as a measure of relative

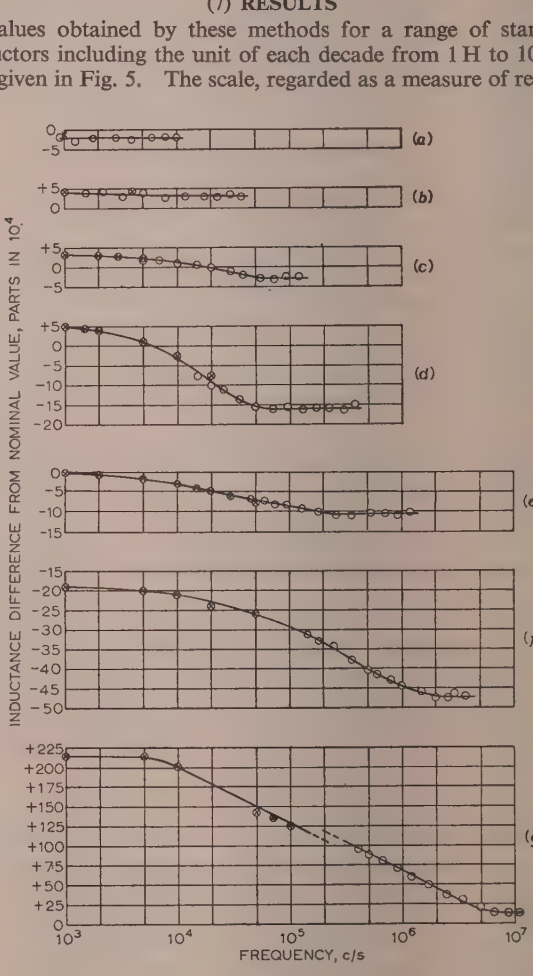


Fig. 5

⊗ Bridge measurements.
○ Resonance method measurements.
(*a*) ES331: 1 H; $C_L = 32{\cdot}16\,\mu\mu$F.
(*b*) ES441: 100 mH; $C_L = 17{\cdot}88\,\mu\mu$F.
(*c*) ES438: 10 mH; $C_L = 15{\cdot}98\,\mu\mu$F.
(*d*) ES435: 1 mH; $C_L = 10{\cdot}97\,\mu\mu$F.
(*e*) ES432: 100 μH; $C_L = 6{\cdot}34\,\mu\mu$F.
(*f*) ES429: 10 μH; $C_L = 5{\cdot}57\,\mu\mu$F.
(*g*) ES426: 1 μH; $C_L = 4{\cdot}2\,\mu\mu$F.

change of inductance with frequency, is the same for all the coils except the smallest (1 μH), for which it is less open in the ratio 5 : 1. The observed values are marked by plain circles, the diameters of which represent values within $\pm 1 \times 10^{-4}$ of the observed value and thus indicate approximately the limits of error that may be expected to arise from the observations of capacitance as a function of frequency of resonance. Ignoring for the moment the smallest coil, it will be seen that within the above limits of error the values all lie on smooth curves of the form that must be expected from considerations of skin effect. For the 1 μH coil the scattering of points about the continuous curve indicates uncertainties of the order of 5×10^{-4} in relative values or $\pm 0{\cdot}000\,5\,\mu$H in absolute value. The uncertainties in the determination of the unwanted residual inductance have been shown to be, at best, of the order of $0{\cdot}000\,2\,\mu$H, and when it is remembered that such inductances cannot be strictly additive it becomes clear that the errors in values obtained by a difference procedure must be expected to be at least of the order of $\pm 0{\cdot}000\,4\,\mu$H in the absolute value, however small the inductance to be measured may be. Thus the best estimate that can be made of the accuracy of the method is $\pm 1 \times 10^{-4} \pm 0{\cdot}000\,4\,\mu$H, since both the sources of error considered above will always be operative. It is very satisfactory that all the experimental values covering so wide a range of conditions should be self-consistent within these limits.

The points on the curves marked by crosses within circles represent values for the same coils measured by bridge techniques by Mr. G. H. Rayner. The curves show clearly that the whole system of measurements is self-consistent to within $2 \times 10^{-4} \pm 0{\cdot}001\,\mu$H. The measurements have been made at intervals over a few years, and some of the discrepancies between low- and high-frequency results probably represent real changes in the coils, for it has recently been observed[3] that they change slightly in value with changing humidity—a variable that was not controlled or recorded in detail when these measurements were made.

The details of the bridge methods are outside the scope of the paper, but since the two techniques are not completely independent, quite apart from the fact that they necessarily start from the same primary standard of inductance, it is desirable to indicate briefly the relation between them. The Campbell primary standard[4] is first used to calibrate an inductometer, or working standard of mutual inductance, at a frequency sufficiently low to ensure that the effects of eddy currents and capacitance are negligible in both standards. The inductometer[5] is then used for the measurement of self-inductance by means of a variant of the Campbell–Heaviside bridge[6] at any frequency low enough to make the effects of capacitance and eddy currents in the inductometer negligible. Such frequencies are well above the limit of 10 c/s imposed by the primary standard, and the measurement gives values of self-inductance at frequencies up to about 100 c/s. Values at higher frequencies are then obtained by means of the Maxwell bridge, in which the self-inductance is balanced against a capacitance. Such a bridge can be balanced with precision at all frequencies up to some tens of kilocycles per second, and by including in the capacitance-arm air capacitors only, for which the variation of capacitance with frequency is negligible, the variation of the inductance with frequency is determined. Having measured the 'frequency correction' for one coil in this way, corresponding values for other coils and for the inductometer are obtained by including the calibrated coil and the standard to be investigated in an appropriate bridge circuit and observing the change of balance point with frequency. The changes in value thus measured necessarily include the effects of both capacitance and eddy currents, and thus in order to obtain the values of inductance for Fig. 5 it is necessary to deduct the effect of self-capacitance which must be separately determined. The value may be known from the radio-frequency calibration, but although it is independent of frequency for any well-designed standard, it necessarily varies with the leads used and with the extent to which capacitance to earth affects the measured quantity. Thus when the bridge circuit includes a Wagner earth the measured change does not include the effects of earth capacitance, while in the resonance method described earlier they are included and contribute to the measured self-capacitance an amount which will depend on whether one terminal of the coil or the other or neither is earth-connected. Such differences in the operative self-capacitance must always be borne in mind when using any standard inductor. They may be estimated by observing the frequency of resonance of the coil in the various conditions, and thus the value appropriate to a bridge measurement with given leads can be deduced. For most of the coils referred to in Fig. 5 the value for bridge measurements with a Wagner earth was some 2 or 3 $\mu\mu$F smaller than that quoted for the resonance method. At the low frequencies the effects of self-capacitance are relatively small, and high accuracy in the value of self-capacitance is usually unnecessary.

To sum up: the low-frequency values of inductance depend primarily on a direct measurement of inductance in terms of the inductometer, and secondly on the measurement of a change with frequency measured by using an air capacitor as a transfer standard of constant capacitance from low to intermediate frequencies. The actual values of the capacitance are not used in these measurements, but merely the fact that they are independent of frequency. The radio-frequency measurements, on the other hand, depend primarily on the actual values of capacitance, which are measured in terms of the inductometer by a variant of the Carey–Foster bridge[6] at frequencies low enough to enable the frequency correction of the inductometer to be deduced from the change of balance point with frequency. These values of capacitance are then assumed to hold good at radio frequencies apart from the effects of residual inductance, which are measured in the way already described.

Several of the methods included in this scheme of measurements have long been practised. An accuracy of 1 part in 10^4 for some of the low-frequency measurements has been obtainable for many years, and a considerably better accuracy is now obtainable in favourable cases. Similarly, it is well known that differences of inductance are detectable without difficulty in radio-frequency circuits with a discrimination closer than the limits claimed above. Nevertheless, it has been the common experience that absolute values obtained by differing techniques often showed inconsistencies and it was impossible to decide which of the assumptions involved was erroneous. It is therefore a matter of some importance that the gap between the radio- and low-frequency techniques has now been closed and that the accuracy of the whole system practised at the N.P.L. has been established within an estimated experimental error that meets all practical requirements.

(8) ACKNOWLEDGMENT

The work here described forms part of the research programme of the National Physical Laboratory and is published by permission of the Director. Several present and former members of the staff of the Electricity Division have from time to time made contributions to it; in particular, it should be recorded that resonance measurements of the requisite sensitivity were first made some 15 years ago by Mr. W. Wilson, and that the experimental arrangements used for the present work were, apart from minor details, among those that he devised.

(9) REFERENCES

(1) O'RAHILLY, A.: 'A Note on Self-Induction', *Journal I.E.E.*, 1940, **86**, p. 179.

(2) AWBERY, J. H.: 'A Simple Method of Fitting a Straight Line to a Series of Observations', *Proceedings of the Physical Society*, 1929, **41**, p. 384.

(3) RAYNER, G. H., and FORD, L. H.: 'The Effect of Humidity on the Stability of Inductance Standards', *Journal of Scientific Instruments*, 1956, **33**, p. 75.

(4) CAMPBELL, A.: *Proceedings of the Royal Society*, A, 1907, **79**, p. 428 and 1912, **87**, p. 391.

(5) ASTBURY, N. F., and FORD, L. H.: 'A Screened Sub-Standard Inductometer', *Philosophical Magazine*, 1938, **25**, p. 1009.

(6) ASTBURY, N. F. and FORD, L. H.: 'The Precision Measurement of Capacitance', *Proceedings of the Physical Society*, 1939, **51**, p. 37.

DISCUSSION ON THE ABOVE PAPER

Dr. J. Brown (*communicated*): One very interesting point in the paper is the discussion in Section 4 showing the validity of lumping the self-capacitance of the inductance into a single element. An alternative approach is to use the Foster expansion for the reactance of any loss-free circuit component, which shows that a representation of the form shown in Fig. A is always

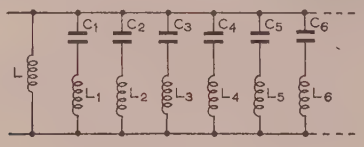


Fig. A

valid. The number of tuned circuits required is, in general, infinite, but if attention is restricted to the frequency range well below the lowest frequency at which any branch resonates, each branch can be approximated by a capacitance. The self-capacitance of the coil is then formed by adding the individual branch capacitances together, giving the required equivalent circuit of the inductance, L, shunted by a single capacitance.

The authors' derivation of this result shows that it can be obtained by an argument more closely linked to the actual physical behaviour of the circuit. It appears that this approach can be simplified as follows: divide the inductance into n sections, each with its self-capacitance as in the paper, giving the representation shown in Fig. B(i). The inductances and capacitances can then be considered to form separate circuit paths, as shown in Fig. B(ii), provided that

$$\ldots i_n = i_{n+1} = \ldots = i \quad . \quad . \quad . \quad . \quad . \quad (A)$$

and

$$\ldots i'_n = i'_{n+1} = \ldots = i' \quad . \quad . \quad . \quad . \quad . \quad (B)$$

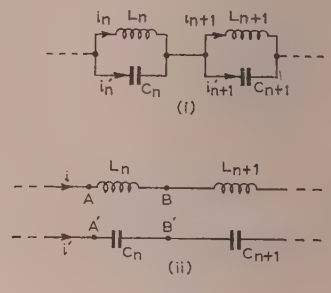


Fig. B

Corresponding points, such as A and A′, B and B′, must be at the same potential for the equivalence between (i) and (ii) in Fig. B to hold, so that

$$\omega L_n i = \frac{i'}{\omega C_n} \text{ for all values of } n \quad . \quad . \quad . \quad (C)$$

i.e. $\omega^2 L_n C_n = i'/i$, which is independent of n. This shows that the products $L_n C_n$ must be a constant, as stated in the paper.

Dr. L. Hartshorn and **Mr. J. J. Denton** (*in reply*): The alternative argument is instructive both for its neatness and the clear picture it presents of the important case of the single-layer solenoid. In this case the measured self-capacitance can clearly be visualized as the turn-to-turn capacitances all connected in series, and experiment, so far as it goes, bears out this notion. Most of our measurements, however, have been made on coils of the multi-layer type, and the argument in the paper was therefore presented in a form which was suggested by the consideration of such coils, although the more complicated relation between self-capacitance and the capacitances between turns has not been explored in further detail.

621.315.5: 621.316.8: 669.018: 389.16

Paper No. 2084 M
July 1956

NICKEL–CHROMIUM–ALUMINIUM–COPPER RESISTANCE WIRE

By A. H. M. ARNOLD, Ph.D., D.Eng., Associate Member.

(*The paper was first received* 29*th October*, 1955, *and in revised form* 22*nd February*, 1956.)

SUMMARY

A review is given of the principal materials used for the construction of resistance standards. The difficulty of producing manganin and constantan commercially with the requisite small value of temperature coefficient at room temperature makes attractive the newer alloys, whose temperature coefficients can be controlled by simple heat treatment. One of these alloys, having the additional advantage of a resistivity three times that of manganin, has been studied at the National Physical Laboratory. It is composed of nickel, chromium, aluminium and copper and is known commercially as Evanohm. When the temperature coefficient, at a given temperature, has been reduced to zero by heat treatment, the curvature of the resistance/temperature characteristic is only one-tenth that of manganin. The stability of resistors constructed of this material has been investigated and has been generally found to be of the order of a few parts in 10^5 per year. The investigations are continuing and it is hoped that better figures may be obtained for well-aged standards. The stability is not adversely affected—and may be improved—by operation at temperatures up to 120° C. Above 140° C there is usually an increase of resistance, but even at 400° C this increase is not rapid. Operation above 400° C is not recommended even for low-accuracy resistors.

Further advantages of the material are a low thermal e.m.f. to copper, high mechanical strength and high ductility. The disadvantages are the necessity of hard soldering and the susceptibility of the resistance to change due to cold working, including vibration.

The resistance standards under investigation for long-term stability are five 1-ohm standards, and one standard of each of the following values: 1 000 ohms, 100 000 ohms, 1 megohm and 10 megohms. An additional 100 000-ohm standard of an alternative design has recently been constructed.

(1) INTRODUCTION

The qualities required of a resistance alloy have been well stated in a recent paper,[1] which incidentally contains a useful list of earlier references. These qualities are:

(*a*) It should have a low temperature-coefficient of electrical resistance over a wide temperature range.

(*b*) It should have a low thermal e.m.f. against copper.

(*c*) Its resistance should be stable over long periods of time.

(*d*) It should be workable by conventional wire and rolling-mill practices.

(*e*) It should be capable of being easily soldered or welded.

To these qualities may be added the following for an alloy required for resistors of high ohmic value:

(*f*) It should be possible to draw it down to wire of 0·001 in or less in diameter.

(*g*) It should have a high resistivity.

(1.1) Resistance/Temperature Characteristic of Resistance Alloys

The resistance of conductors, whether metals or alloys, as distinct from semi-conductors, is lower at the absolute zero of temperature than at the melting temperature. The average temperature-coefficient over this range is therefore positive. The ideal of a resistance alloy with zero temperature-coefficient over the whole range from absolute zero to melting temperature does not seem likely to be attainable, and investigators have concentrated their attention on developing alloys with a change of curvature in their resistance/temperature characteristics, so that over a restricted temperature range the temperature coefficient may be very small. Fig. 1 shows the ideal type of curve, while

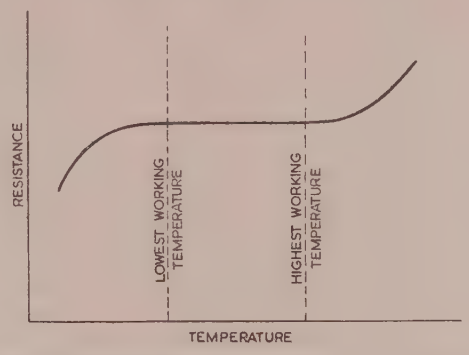


Fig. 1.—Ideal resistance/temperature curve of an alloy.

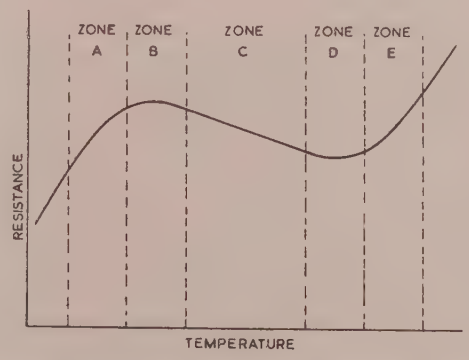


Fig. 2.—Practical approximation to ideal curve of Fig. 1.

Fig. 2 shows the sort of curve which is more likely to be attained in practice.

In Fig. 2 the part of the resistance/temperature curve which is of interest has been divided into five zones: zone A of positive temperature-coefficient at temperatures below that at which the resistance reaches a maximum, zone B containing the maximum value, zone D containing a minimum value, zone C intermediate between zone B and zone D and having a negative temperature-coefficient and zone E of positive temperature-coefficient at temperatures above that at which there is a minimum resistance value. If an alloy can be produced in which the maximum and minimum resistances are very nearly equal, resistors of this alloy may be operated over the whole temperature range of zones B, C and D. If, however, the maximum and minimum resistances differ appreciably, the operating temperature range should be restricted to zone B or zone D. The width of zone C shown in Fig. 2 may vary from a few degrees to hundreds of

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.
The paper is an official communication from the National Physical Laboratory.

degrees, according to the alloy. The resistance/temperature relation may also be more complicated than that shown.

(1.2) Early Resistance Alloys

The earliest resistance alloy used was probably the copper–nickel–zinc alloy known as german silver. This alloy had a positive temperature-coefficient of about one-tenth that of copper and the zinc content was a source of instability. A better alloy was the binary alloy of $\frac{2}{3}$ silver and $\frac{1}{3}$ platinum which was used in the middle of the last century for the construction of the British Association resistance standards. It was not realized at that time that a resistance/temperature curve of the shape of Fig. 2 was attainable, and it was not until about 1888 that the first alloy with a negative temperature-coefficient at room temperature was discovered. The discovery seems to have been made by Edward Weston, an Englishman working in the United States, but there was much parallel work in Germany at the Physikalisch-Technischen-Reichsanstalt (P-T-R) and at the Isabellen Metal Works (now the Isabellen-Hutte Heuser K.G.). The alloy was the binary alloy of 60% copper and 40% nickel described by the Germans as Konstantin and more commonly known in England as constantan. Since that date the composition of constantan and other similar alloys has varied somewhat, so that room temperature has occurred variously in zone A, zone B or zone C. It has never proved commercially practicable to produce constantan of such a kind that room temperature was consistently in zone B, and the best wire has only been obtained by selection. Constantan has been extensively used for a.c. resistors required to operate over considerable temperature ranges, but it is less satisfactory for d.c. resistors on account of its high thermal e.m.f. to copper, about 40 μV per deg C. For use with direct current it has been almost entirely superseded by another alloy discovered a few years later. This was the ternary alloy of copper, manganese and nickel which was given the name of manganin by the Germans. This name is now almost universally used, although other proprietary names have been introduced by various manufacturers. It is possible that this alloy was also discovered by Edward Weston, but this is not certain. The composition varied appreciably at first; the modern composition of approximately 84% copper, 12% manganese and 4% nickel probably dates from about 1895.

Manganin has a resistivity of about 40 microhm-cm and a thermal e.m.f. to copper of 2 μV per deg C. Room temperature may occur in zone A, zone B or zone C, but it has proved somewhat easier to obtain material such that room temperature lies in zone B than it was with constantan. In two respects manganin is inferior to constantan. First, it is more difficult to solder. Soft soldered joints are not only difficult to make but are often a source of instability of resistance. Secondly, the curvature of the resistance/temperature curve is greater than that of constantan, so that even though the mean operating temperature of a resistance standard is exactly that at which the resistance is a maximum, there is a second-order resistance change with temperature which is greater for manganin than for constantan. However, the advantage of a low thermal e.m.f. to copper outweighs these disadvantages. In respect of stability of resistance it is quite possible that constantan is superior to manganin since it is less subject to surface oxidation. It must, however, be remembered that the stability of a resistance standard is due only in part to the qualities of the resistance alloy and is dependent to a great extent on the design and construction of the standard. The instant popularity of manganin for resistance standards must be largely ascribed to the fine workmanship of Otto Wolff, who specialized in the manufacture of the Reichsanstalt resistors in universal use as national standards until well after the First World War. Since 1914 a great deal of work has been carried out both here and in the United States on improving the method of construction of resistance standards, and both countries now possess a number of 1-ohm manganin standards which change their values by less than one part in a million each year. The development work necessary for the production of these standards has occupied a period of over fifty years, and an equal or greater period may well elapse before the stability of resistance standards of newer alloys is established with sufficient certainty to enable them to compete with manganin. Although manganin is well established for national standards of resistance, it has not been possible to achieve commercial production of manganin to the same consistent high standard. Some manganin is unstable in resistance, and large quantities are produced of a kind in which room temperature is in zone A or zone C. Manganin joints have also been a source of weakness. Thus there may well be room in the commercial production of resistors for an alternative alloy, even though manganin is now too well established for national standards to be in danger of displacement for many years.

Before this aspect is considered, however, the efforts which have been made to develop alternative materials for national standards will be briefly reviewed.

(1.3) Alternative Materials for National Resistance Standards

Resistors have been constructed at the National Physical Laboratory with platinum as the resistance material in the belief that it is less subject to physical or chemical change. To obtain the same accuracy with these standards as was obtained with the manganin standards it was necessary to maintain the temperature of the coils constant to one four-thousandth of a degree centigrade or better. The difficulties involved in such close temperature control were largely overcome, but the experiment was ultimately suspended when the results showed that the superiority of platinum over manganin with respect to stability could not be established with any certainty.

At the National Bureau of Standards in the United States, J. L. Thomas investigated the properties of gold–chromium alloys, and was successful in producing samples in which the width of zones B, C and D together was only about 10° C and the difference between the maximum resistance in zone B and the minimum resistance in zone D was less than one part in a million. Moreover, these zones could be brought to the neighbourhood of room temperature by heat treatment at the low temperature of 150° C. The stability of resistors constructed with this material, however, appears to be inferior to that of manganin resistors and the thermal e.m.f. to copper is three or four times as great.

Thomas also investigated the properties of an alloy which had been developed commercially as early as 1910. This alloy, known as Therlo, differs in composition from manganin mainly in the replacement of the nickel content by aluminium. Thomas made a series of alloys of slightly different composition and found the best to be 85% copper, 9·5% manganese, 5·5% aluminium and a very small amount of iron. The resistivity of this alloy was similar to that of manganin, while the thermo-electric e.m.f. to copper was only about one-tenth. The curvature of the resistance/temperature curve at the maximum resistance value in zone B was about one-half that of manganin. The most remarkable property of this alloy, however, was that the temperature for maximum resistance in zone B could be altered by heat treatment at the low temperature of 140° C. Provided that suitable insulation was used it was therefore possible to adjust the temperature coefficient of a completed resistance standard to zero at the mean working temperature. Early results showed

that resistance standards of Therlo had a stability comparable to those of manganin, and if these results should be confirmed over the years it would appear that Therlo should be a powerful competitor to manganin. Alfred Schulze, who carried out work along the same lines in Germany during 1933–41, was unable to produce stable resistors of Therlo but succeeded with alloys having similar qualities but slightly different compositions. Two of these alloys are known by the commercial names of Isabellin and Novokonstant.

The present position, therefore, is that the three alloys, Therlo, Isabellin and Novokonstant, are superior to manganin in respect of their thermal e.m.f. to copper and may have a slight superiority in respect of the curvature of the resistance/temperature curve in zone B. They have the great advantage over manganin that the mid-temperature of zone B can be controlled by heat treatment at a temperature considerably lower than the annealing temperature, and lower than the maximum operating temperature of some insulating materials. Their stability over long periods is not yet established and they are likely to be unsuitable for operations involving large temperature rises.

(1.4) High-Resistivity Alloys

None of these alloys meets the demand for a material of high resistivity, and until recent years the only alloy available was the binary alloy of 80% nickel and 20% chromium which has a resistivity at room temperature about three times that of manganin and a positive temperature-coefficient of about 60 parts in 10^6 per deg C. The maximum resistance in zone B occurs at about 500° C. The temperature coefficient is rather high for the better classes of resistors and the stability of resistance is not good. The curvature of the resistance/temperature relation in zone A is very small.

The stability of the nickel–chromium alloy has been greatly improved, and the valuable property of control of temperature-coefficient by heat treatment at moderate temperatures has been achieved by the addition of two more metals. Two alloys of this type are of particular interest. The one, commercially known as Evanohm, has an approximate composition of 73% nickel, 21% chromium, 2% aluminium, 2% copper and the balance of other metals. The other, commercially known as Karma, has a similar composition except that the copper is replaced by iron. Both these alloys have a resistivity three times that of manganin, a thermal e.m.f. to copper generally rather less than that of manganin, high tensile strength (making possible the drawing of fine wires) and a temperature-coefficient which may be controlled by heat treatment at moderate temperatures. Moreover, when the temperature-coefficient is adjusted so that the mean working temperature comes in the middle of zone B, the curvature of the resistance/temperature relation is about one-tenth that of manganin and less than one-half that of constantan. The alloys must be hard-soldered. Apart from this one disadvantage, they appear admirably suited for resistors of high value or required to operate over large temperature ranges. Investigations of their stability have been started both at the National Bureau of Standards and at the National Physical Laboratory.

The results obtained at the National Physical Laboratory so far have shown that the alloy containing copper is superior in stability to the one containing iron. Results of tests on the copper alloy only will be given in the paper. The samples tested were obtained commercially, and it does not necessarily follow that all samples of the copper alloy will be superior to samples of the iron alloy, especially in view of the conflicting results which have been obtained by other workers on other alloys. However, the purpose of the investigation was to determine the suitability of one of these alloys for resistance standards, and the results obtained so far have been encouraging.

(2) THE TEMPERATURE-COEFFICIENT OF ELECTRICAL RESISTANCE OF EVANOHM

Experiments were carried out at the National Physical Laboratory to confirm and amplify the information given in an earlier paper[2] on the resistance/temperature characteristic.

The annealing temperature of Evanohm is in the neighbourhood, of 1 000° C, and in wire form the alloy may be annealed either by heating in an oven at this temperature or by what is sometimes a simpler process—that of passing sufficient current through the wire to raise it to the required temperature. In either case the cooling must be sufficiently rapid to prevent changes in resistance and temperature-coefficient occurring at lower temperatures. It was found that when annealing was carried out in air the wire was slightly tarnished by oxidation. Some samples were annealed in hydrogen and in a partial vacuum in the hope of avoiding oxidation, but the experiment was abandoned when it was found that the temperature coefficient was about 50% higher than when annealing was carried out in air, and that it could not be reduced by subsequent heat treatment at a lower temperature. When properly annealed the wire has a temperature-coefficient of electrical resistance similar to that of the nickel–chromium alloy, namely approximately 60 parts in 10^6 per deg C with negligible variation over a temperature range from room temperature or lower to about 400° C. It is not possible to determine a true value of the temperature-coefficient above 400° C, since part of the resistance change is permanent above this temperature. This permanent resistance change is at first an increase and is accompanied by a decrease of temperature coefficient at room temperature. Both the amount and rate of change are dependent on the temperature. With heat treatment at 550° C the resistance first increases approximately 15% and then decreases. The temperature coefficient first decreases through zero to a negative value of about 20 parts in 10^6 per deg C and then increases to a positive value approaching that for the annealed condition. However, the resistance/temperature curve loses its linearity after prolonged heat treatment, and it is not desirable to continue the treatment beyond the point at which the temperature-coefficient first becomes zero.

There is some difficulty in achieving the correct amount of heat treatment, since in the neighbourhood of zero temperature-coefficient the wire may change its characteristics very rapidly. Commercial heat treatment is controlled to give a temperature-coefficient within the limits of ± 20 parts in 10^6 per deg C, but commercial supplies within the limits of ± 5 parts in 10^6 can be obtained by selection. The temperature coefficient may be affected to some extent by manufacturing processes carried out after heat treatment, such as insulating the wire. When, therefore, it is desired to construct a resistance standard with zero temperature-coefficient at a particular temperature, the final heat treatment must be carried out after the completion of all processes involving cold working of the wire.

The linearity of the resistance/temperature relation of the annealed wire is not maintained as the temperature-coefficient is reduced by heat treatment. The curvature introduced is very small and appears to be dependent to some extent on the temperature at which heat treatment is given. When the temperature-coefficient is zero the curvature is about one-tenth that of manganin so that the working temperature range of a resistor may be three times as great for the same change of resistance.

(3) THE CAUSES OF INSTABILITY OF RESISTANCE

Resistance changes in the alloy itself may be due to physical or chemical changes occurring either spontaneously or as a result of treatment of the wire.

Chemical changes occur mainly on the surface of the wire and

may be accelerated by heat, humidity and other causes or retarded by protective coatings or by immersion in a dry atmosphere of an inert gas. Insulating liquids have also been used to protect the surface of wires, but there is more difficulty in obtaining an inert liquid than an inert gas.

Physical changes occur mainly in the body of the wire and may be accelerated by heat treatment or by cold working, including vibration.

The importance of surface changes becomes greater as the ratio of surface area to volume is increased, and it is therefore an advantage to use a wire as large in diameter as practicable. In this respect Evanohm has an advantage over manganin for resistances of all values on account of its higher resistivity. It seems probable also that the surface is less liable to contamination than that of manganin, and the encouraging results of stability tests on resistors constructed with wire of 0·001 in diameter lend support to this view. These results are given in Sections 6–10.

In respect of physical changes it is not so likely that Evanohm will show up advantageously. The presence of aluminium in its composition is believed to be favourable to enlargement of grain size with time, but it is also believed that the presence of copper will counteract this tendency. However this may be, it is certain that both the resistance and the temperature-coefficient of this alloy are peculiarly susceptible to cold working in any form, including vibration. Further, the very advantage of being able to control the temperature-coefficient by heat treatment at moderate temperatures makes the alloy suspect if it is normally operated over a wide temperature range.

It is unlikely that Evanohm will prove a formidable competitor to manganin for resistance standards of low value and of the highest class, since, on the one hand, the advantage of a vanishingly small temperature-coefficient is of minor importance for resistance standards which are operated under closely controlled temperature conditions and with small self-heating, and, on the other hand, manganin standard resistances of proved stability are already in existence, whereas Evanohm is unproved. The field for Evanohm is more likely to be in working resistance standards which are operated over a considerable temperature range on account of self-heating. The exact temperature of a resistor heated by its own current is always difficult to determine, so that it is a great advantage if it is constructed of wire of small temperature coefficient, while a moderate amount of secular instability may be relatively unimportant for a working standard. Working resistance standards of manganin are quite commonly operated over a temperature rise of 30° C, so that for the same performance it should be possible to operate an Evanohm standard over a range of 100° C. This would make possible the construction of a standard of smaller physical dimensions at lower cost. For a.c. work the smaller dimensions would generally result in a reduction of errors from capacitance. It is therefore of importance to establish the stability of Evanohm when subjected to cyclic temperature changes of at least 100° C, and it may be observed that, although a slow drift in resistance might not matter in a working standard, a corresponding drift in the temperature-coefficient would be objectionable unless it were extremely slow.

The stability of manganin resistance standards of high value—1 000 ohms and upwards—is not so good as those of lower value, and Evanohm may prove competitive, even for standards of the highest class, in this field.

The experimental work carried out to determine the stability of Evanohm is described in Sections 5–10. Much of it required measurements of the highest precision on properly constructed standards, but in order to accelerate the work many measurements were made to a lower order of accuracy on resistors of simple construction. When the preliminary results proved encouraging, a number of properly made standards were constructed for long-term measurements of the highest accuracy.

(4) EFFECT OF HEAT TREATMENT

The simplest method of applying heat treatment to short lengths of Evanohm wire is to pass current through them. It was found experimentally that a current of 1·2 amp through a wire of 0·004 8 in diameter supported horizontally in air was sufficient to raise it to the annealing temperature of 1 000° C. Maintenance of this current for five to ten minutes restored the wire to the annealed condition, whatever the previous heat treatment at lower temperatures. The annealed wire has a temperature-coefficient of +60–65 parts in 10^6 per deg C at 20° C.

A number of samples of annealed wire were subjected to currents between 0·85 amp and 1·2 amp for varying periods from a few seconds to five minutes, and the temperature-coefficient at 20° C was determined after the heat treatment. The values obtained were all within the range of +55–70 parts in 10^6, and it was concluded that the condition of the wires did not differ significantly from the annealed condition. When the current was reduced below 0·7 amp its duration had a marked effect on the final temperature-coefficient, and Fig. 3 shows the results

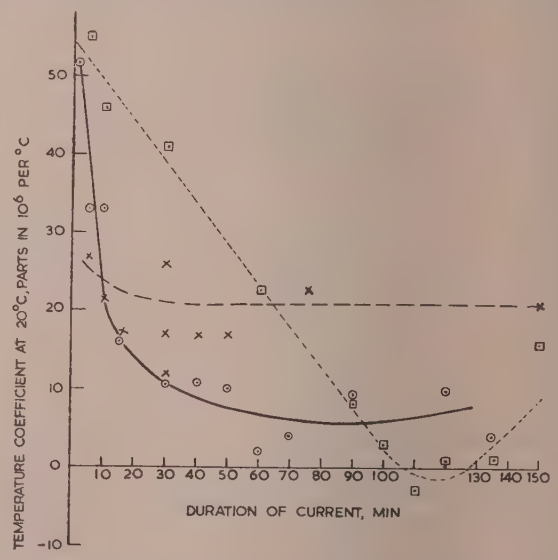


Fig. 3.—Effect on temperature-coefficient of passing current through Evanohm wire supported horizontally in air.

Wire diameter 0·004 8 in.
0·60 amp.
0·65 amp.
0·68 amp.

obtained with currents of 0·68, 0·65 and 0·60 amp. There are inconsistencies in many of the observed points but the general trend is clear: the temperature-coefficient is reduced by the passage of the current. The rate of reduction falls as the current is reduced although the maximum change of coefficient increases and is eventually sufficient to reverse its sign. The temperature of the wire with a current of 0·6 amp passing through it is probably between 500 and 600° C.

An alternative method of heat treatment is to place the sample in a heated oven. It is easier by this method to maintain the maximum temperature of the sample at a constant and known value, but as the oven takes some hours to warm up, and even longer to cool down, part of the heat treatment of the wire occurs at a temperature below the maximum value. A sample of

annealed 0·004 8 in-diameter wire was subjected to prolonged heat treatment in an oven according to the following procedure.

The oven was warmed up slowly from room temperature, and the resistance/temperature relation of the wire was determined up to 300° C. A final temperature of 450° C was reached in 2½ h and the oven was then allowed to cool off to room temperature. On succeeding days the maximum temperature was raised by 25° C until a maximum of 550°C was reached. On all succeeding days the maximum was 550° C, but on some days it was maintained for a period before the oven was allowed to cool. It was found that over the whole temperature range from 20° C to 300° C the resistance/temperature relation could be represented very closely by the parabolic law given in eqn. (1):

$$R_t = R_{20}[1 + \alpha(t - 20) + \beta(t - 20)^2] \quad . \quad . \quad (1)$$

where R_t = Resistance at a temperature of t deg C.
R_{20} = Resistance at a temperature of 20° C.
α and β are constant coefficients.

Table 1 shows the experimentally determined values of α and β.

For any given experimental resistance/temperature curve it is possible to vary α, provided that β is also suitably varied, over a small range with very little loss of fit. The values given in Table 1 are those which give the best fit to each individual curve and not those which give the most consistent results for the complete set of curves. The Table shows that the value of α falls from $+58 \times 10^{-6}$ for the annealed wire to a negative maximum of -23×10^{-6}. Thereafter, further heat treatment causes it to change in a positive direction. In another experiment, in which the oven temperature was eventually raised to 700° C in order to accelerate the heat treatment, the average temperature-coefficient over the range 20–100° C finally reached a value of $+84 \times 10^{-6}$, but over the range 20–300° C the average coefficient was only $+54 \times 10^{-6}$. The resistance/temperature relation, although approximately similar to that for the annealed condition, had a much greater curvature and there was no indication that heat treatment at this temperature would ever bring the wire back to the annealed condition.

The variation of the coefficient β with heat treatment follows a somewhat irregular course until the maximum negative value of α is reached. Thereafter β increases fairly consistently in the negative direction. The desirable objective is, of course, to heat-treat the wire so that both α and β are zero simultaneously. It is apparent from Table 1 that this objective is not likely to be achieved when α passes through zero from a negative to a positive value, and in all later work heat treatment was designed to bring the coefficient α to the first zero, changing from a positive to a negative value. The irregular variations of the coefficient β in the early stages of heat treatment suggest that the exact nature of the heat treatment might affect the value of β when α is zero. The correct heat treatment to make α and β simultaneously zero has not yet been found, although a few experiments with different heat treatments were tried.

The third point of interest emerging from Table 1 is that the change of α from its value for the annealed wire is approximately proportional to the change of resistance from its value for the annealed wire. This relation may be expressed by eqn. (2) thus

$$\frac{R_2}{R_1} = 1 + 2000(\alpha_1 - \alpha_2) \quad . \quad . \quad . \quad . \quad . \quad (2)$$

where R_2 is the resistance corresponding to the coefficient α_2 and R_1 is the resistance corresponding to the coefficient α_1. This equation is sometimes convenient when heat-treating a wire to obtain zero temperature coefficient. It is also of value in judging whether a secular resistance change of given amount is likely to be accompanied by a significant change of temperature-coefficient.

It is apparent from the results given in Fig. 3 and Table 1 that the temperature coefficient of the wire may be changed by heat treatment at various temperatures. The rate of change becomes less as the temperature is lowered, but it is important to discover whether significant changes occur at moderate temperatures, since they might limit the possible working temperature of a resistor. It was found possible to bring the temperature coefficient of annealed wire to zero by heat treatment at a temperature of 400° C. The time required for this was 265 h. It would not, therefore, be practicable to operate a precision resistor up to

Table 1

EFFECTS OF PROLONGED HEAT TREATMENT OF 0·004 8 IN-DIAMETER WIRE

Day	Maximum oven temperature	Period at maximum temperature	α	β	Increase in resistance (Y)	Remarks
	deg C	hours			%	
			$+58 \times 10^{-6}$	$+12 \times 10^{-9}$	0·0	Annealed wire Average temperature coefficient over range 20–300° C = $+ 61 \times 10^{-6}$
1	450	Momentary	$+49 \times 10^{-6}$	-10×10^{-9}	2·0	
2	475	Momentary	$+43 \times 10^{-6}$	-23×10^{-9}	3·8	
3	500	Momentary	$+36 \times 10^{-6}$	-32×10^{-9}	5·4	
4	525	Momentary	$+12 \times 10^{-6}$	-21×10^{-9}	8·5	
5	550	Momentary	-11×10^{-6}	$- 8 \times 10^{-9}$	11·3	
6	550	Momentary	-19×10^{-6}	$- 3 \times 10^{-9}$	12·4	
7	550	Momentary	-22×10^{-6}	$- 4 \times 10^{-9}$	13·0	
8	550	0·5	-21×10^{-6}	-12×10^{-9}	13·5	
9	550	1	-23×10^{-6}	-10×10^{-9}	14·1	
10	550	Momentary	-21×10^{-6}	-11×10^{-9}		Voltage leads of copper burnt away. Evanohm leads substituted
11	550	2	-15×10^{-6}	-29×10^{-9}		
12	550	2	-11×10^{-6}	-35×10^{-9}		
13	550	2	$- 8 \times 10^{-6}$	-40×10^{-9}		
14	550	2	$- 5 \times 10^{-6}$	-44×10^{-9}		
15	550	4	0	-51×10^{-9}		
16	550	0·5	$+ 1 \times 10^{-6}$	-53×10^{-9}		
17	550	4	$+ 4 \times 10^{-6}$	-56×10^{-9}		
18	550	4	$+ 7 \times 10^{-6}$	-60×10^{-9}		

400° C. However, if the accuracy required were only 1% it would be possible to adjust the temperature coefficient of the wire to +25 parts in 10^6 per deg C initially, and the coefficient would gradually decrease with time of operation at 400° C, eventually reaching a negative value of about 25 parts in 10^6 per deg C. The coefficient would then begin to change in a positive direction. It is likely that the time of operation at 400° C would exceed 1 000 h before the coefficient returned to its initial value of +25 parts in 10^6 per deg C. It would then be necessary to anneal the resistance material and readjust the temperature coefficient by heat treatment. There would be a resistance change between 20 and 400° C on account of temperature-coefficient of 1% or less. There would also be a resistance change of about 10% accompanying the change of temperature-coefficient. It would, therefore, be necessary to adjust the resistance at intervals in order to maintain it near to its nominal value.

For precision work, changes of this magnitude would not be tolerable, and experiments were carried out to determine what was the maximum possible operating temperature for accurate work. It is not possible to assign a rate of change of resistance and temperature-coefficient to each maximum operating temperature, since the rates of change vary with time of operation. The resistance changes occurring in one resistor with various periods of operation at various temperatures have, however, been ascertained, and are shown in Table 2. The temperature-

Table 2

CHANGE OF RESISTANCE WITH TIME AT VARIOUS TEMPERATURES OF A RESISTOR OF 0·004 8 IN-DIAMETER WIRE

Time	Temperature	Change of resistance at 20° C	Total change of resistance at 20° C
h	°C	Parts in 10^6	Parts in 10^6
1	100	+4	+4
24	100	−36	−32
4	200	+46	+14
24	200	+137	+151
16	250	+169	+320
20	300	+950	+1 270
24	100	+12	+1 282
20	150	+10	+1 292
24	150	+8	+1 300
20	150	−6	+1 294

coefficient at the start of the tests was −4 parts in 10^6 per deg C, and it changed by less than one part in 10^6 per deg C during the tests. It is apparent from this Table that the resistance change at 300° C is not tolerable, and that even at 200° C the changes are larger than are desirable. At 150 and 100° C the changes are small and irregular in direction, and the effect of prolonged operation at a temperature of about 150° C was therefore studied in more detail. A 10 000-ohm coil of 0·001 in-diameter wire was immersed in silicone oil. On five days each week the temperature of the oil was raised to 143° C in a period of 3 h and held at that temperature for 5 h. This treatment was continued for 12 months. During the first three months the resistance increased by 29 parts in 10^6 and after this time it increased at an average rate of 20 parts in 10^6 per year. Thus, for resistors intended to be accurate within one part in 10^4 it would only be necessary after the first three months to adjust the value once every ten years, on account of operation at 143° C. It is, therefore, quite practicable to operate up to 140° C, but because of secular change the maximum working temperature should probably not exceed 120° C on account of the curvature of the resistance/temperature characteristic. From tests on a number of samples it appears that β usually lies between -20×10^{-9} and -40×10^{-9} for small values of α, and if the temperature-coefficient is adjusted so that the maximum resistance occurs at 70° C, the resistance at 20° C and at 120° C is 0·005% less than at 70° C for the lower limits of β and 0·01% less for the upper limit of β.

(5) LONG-TERM STABILITY

Whilst the experiments already described were proceeding the long-term stability of a number of resistors was investigated. It was generally found that the stability, after an initial settling-down period, was of the order of a few parts in 10^5 per year, and was quite adequate for resistors which were to be operated with a high temperature-rise and which could be readjusted every few years to their nominal value. There was little evidence of a definite drift in one direction, except during periods of high-temperature operation, and the resistance variations which occurred might well be due to imperfections in the construction of the resistors rather than defects in the material itself.

For resistors of the highest class, operated over a small temperature range, say from 15 to 25° C, a higher stability was desired and a few well-made resistors were selected for study in this respect. It was considered that Evanohm was likely to be used for high-resistance standards rather than for low-resistance ones and the early coils manufactured were of high value. Standards of lower value have been constructed more recently, but there is little data on stability yet available for these.

The method of construction of standards may affect their stability, and a description of this for each standard is given in Sections 6–10, together with the resistance measurements.

The method of soldering the copper lead-in wires to the resistance wire was the same for all standards. It is somewhat similar to a method already described;[3] joints have been found easy to make and have not subsequently given trouble. The section of a stainless-steel rod, $\frac{1}{4}$ in in diameter and 5 in long, was reduced to about $\frac{1}{40}$ in^2 over the middle 2 in lengths, by grinding two flats. A small bead of a special copper–silver solder was located in a small depression drilled in the middle of one flat. Pure borax was used as flux and the minimum heat necessary was provided by passing an appropriate current, 100 to 150 amp, through the rod. The surface of the Evanohm wire was cleaned by scraping with a sharp-edged tool before soldering. After soldering, the borax on the joint was cracked and the joint immersed and vibrated for a few minutes, first in dilute sulphuric acid, then successively in distilled water, an ammonia solution and finally in distilled water again.

(6) 100 000-OHM STANDARD

The wire used for this standard was 0·001 in in diameter; it was enamelled and had a resistance of 760 ohms/ft. The preparatory heat treatment was applied by the manufacturer before enamelling and no further heat treatment was carried out. The average temperature-coefficient over the range 20–50° C was +5 parts in 10^6 per deg C, the variation between various samples, in general, not exceeding one part in 10^6.

The wire was wound in a single layer on a silk-covered cylindrical metal former and mounted in a case of a type described in an earlier paper,[4] making what is usually known as a Class S standard. The insulation resistance of the coil mounting was 10^{12} ohms. After construction, the coil was allowed to age at room temperature for six months. It was then placed in a desiccator for two years, after which it was finally sealed.

The resistance was measured to 1 part in 10^5 during the period in which it was in the desiccator and to 1 part in 10^5 after sealing.

Table 3

RESISTANCE OF 100 000-OHM S-COIL

Date	Temperature	Resistance	Remarks
	°C	ohms	
11.7.1952	20	100 009	After one week in desiccator
			Insulation resistance 5×10^{10} ohms
8.12.1952	20	100 007	In desiccator
			Insulation resistance $1{\cdot}6 \times 10^{11}$ ohms
22.12.1953	20	100 007	In desiccator
			Insulation resistance $0{\cdot}8 \times 10^{12}$ ohms
6.5.1954	20	100 008	In desiccator
			Insulation resistance $1{\cdot}3 \times 10^{12}$ ohms
7.7.1954	20	100 011	After removal from desiccator and sealing in dry air
			Insulation resistance $0{\cdot}9 \times 10^{12}$ ohms
7.7.1954	20	100 012·5	Measured on precision bridge
23.7.1954	15	100 007·0	
23.7.1954	20	100 012·6	Average temperature-coefficient +0·000 011
23.7.1954	25	100 018·2	
22.9.1954	20	100 012·8	
29.4.1955	20	100 014·6	
24.10.1955	20	100 015·1	
7.2.1956	20	100 015·3	

Table 3 shows the measurements made up to the present time. The resistance fell slightly during the first 5 months in the desiccator and thereafter it increased about 4 parts in 10^5 during the remaining 19 months in the desiccator. Since the coil was sealed the resistance has increased a further $2\frac{1}{2}$ parts in 10^5 in 15 months. The resistance has thus increased at an approximately constant rate of 2 parts in 10^5 per year for the last 3 years. It may also be observed that the temperature-coefficient of the standard is about twice that of the samples of wire tested before it was constructed. This difference might be due to differences in the temperature-coefficient of the wire along its length, but, in view of the later evidence showing a steady secular resistance increase, it appears more likely that both the resistance increase and the higher temperature-coefficient arise from strains imposed on the wire by the former on which it is wound.

The 1-megohm and 10-megohm standards, described in Sections 7 and 8, were constructed on a different principle, the wire being wound on mica cards. The temperature-coefficient of these standards was found to be substantially the same as that of samples of the wire used for them; it is therefore probable that the strain involved in this method of winding is small. A new 100 000-ohm standard has been constructed in which the wire is wound on a mica card. This standard is mounted in a rectangular metal box with S-type terminals. The temperature-coefficient of the completed standard is +4 parts in 10^6 per deg C over the range 15–25° C, substantially in agreement with the value obtained for samples of the wire used. The resistance was 99 969·5 ohms on 24th October, 1955 (immediately after sealing), and 99 969·8 ohms on 7th February, 1956.

(7) SUBDIVIDED 1-MEGOHM RESISTANCE STANDARD

This resistor was constructed with 0·001 in-diameter enamelled Evanohm wire similar to that used for the 100 000-ohm S-coil. The wire was wound in a single layer on thin mica cards, each card carrying a length measuring 100 000 ohms. The cards were mounted in an unsealed Perspex box. The ends of the wire and tappings for each 100 000 ohms were brought to terminals on the top of the box. Table 4 shows the resistance values obtained.

The temperature-coefficient of the completed standard was +4 parts in 10^6 per deg C over the range 15–25° C. This value is substantially the same as that of the wire before winding.

Table 4

SUBDIVIDED 1-MEGOHM RESISTANCE STANDARD

Sections	Resistance at 20° C at dates given				
	15.3.54	10.8.54	4.4.55	4.8.55	2.2.56
	ohms	ohms	ohms	ohms	ohms
1	100 003	100 006	100 004	100 006	100 006
1 and 2	200 005	200 012	200 008	200 011	200 011
1–3	299 997	300 008	300 001	300 006	300 006
1–4	400 008	400 023	400 013	400 019	400 020
1–5	500 015	500 033	500 021	500 028	500 029
1–6	599 998	600 020	600 006	600 014	600 015
1–7	699 965	699 991	699 975	699 984	699 985
1–8	800 007	800 036	800 017	800 027	800 028
1–9	900 048	900 080	900 058	900 069	900 070
1–10	1 000 003	1 000 039	1 000 014	1 000 027	1 000 028

(8) SUBDIVIDED 10-MEGOHM RESISTANCE STANDARD

The construction of this standard was generally similar to that of the subdivided 1-megohm standard, except that each card contained wire measuring 500 000 ohms and the whole resistor was subdivided into ten sections of 1 megohm each. Table 5 shows the resistance values obtained.

Table 5

SUBDIVIDED 10-MEGOHM RESISTANCE STANDARD

Sections	Resistance at 20° C at dates given			
	1.12.54	5.4.55	13.8.55	2.2.56
	megohms	megohms	megohms	megohms
1	1·000 02	1·000 02	1·000 03	0·999 93
1 and 2	2·000 06	2·000 07	2·000 10	1·999 98
1–3	3·000 08	3·000 09	3·000 13	3·000 01
1–4	4·000 07	4·000 09	4·000 14	4·000 02
1–5	5·000 06	5·000 08	5·000 15	5·000 03
1–6	6·000 09	6·000 12	6·000 20	6·000 08
1–7	7·000 12	7·000 15	7·000 24	7·000 12
1–8	8·000 09	8·000 19	8·000 30	8·000 18
1–9	9·000 25	9·000 35	9·000 47	9·000 34
1–10	10·000 19	10·000 28	10·000 41	10·000 28

The temperature-coefficient of samples of the wire used was found to be within the limits of ± 1 part in 10^6 per deg C, while that of the completed standard was -2 parts in 10^6 per deg C over the range 15–25° C.

(9) 1 000-OHM STANDARDS

Three 1 000-ohm lengths of wire wound on mica cards were subjected to various kinds of heat treatment, to ascertain whether the stability was affected. The wire used was 0·004 8 in in diameter and it had been enamelled by the manufacturers after adjustment of the temperature-coefficient to a low value.

The first coil was maintained at air temperature after construction. Its resistance increased 6 parts in 10^6 during the first 3 months and then decreased 8 parts in 10^6 during the next 4 months; thereafter, it decreased at a constant rate of 13 parts in 10^6 per year for 1 year.

The second coil was maintained at 200° C for 25 h and then allowed to rest for one week before measurements were made. Its resistance has remained constant to within ± 3 parts in 10^6 for 20 months without showing any steady drift.

The third coil was maintained at 200° C for 35 h, during which time its resistance increased by 227 parts in 10^6. This was followed by a period of 30 h at 180° C entailing an increase of resistance of 17 parts in 10^6. A further period of 155 h at 160° C caused an increase of 34 parts in 10^6.

The temperature was then reduced to and held at 140° C. Measurements at long intervals gave the following results:

After the first 500 h at 140° C resistance increased 10 parts in 10^6.

After the second 500 h at 140° C resistance decreased 25 parts in 10^6.

After the third 500 h at 140°C resistance unchanged.

After the fourth 500 h at 140° C resistance increased 2 parts in 10^6.

The coil was then maintained at air temperature for 3 months, during which time its resistance increased by 4 parts in 10^6.

These figures, although not conclusive, suggest that some heat treatment at moderate temperatures is beneficial. Some strain is always put on the wire during the construction of a resistor, and the heat treatment probably removes it. The final stability figures of a few parts in 10^6 for the second and third coils were very good, and justified the construction of an S-type 1 000-ohm coil for further investigation. The coiled-coil method of construction[5] was adopted and the wire was annealed after coiling to remove strains.

The wire used was 0·004 8 in in diameter and was coated with synthetic enamel. Two processes for removing the enamel were tried. In the first the wire was immersed in pure phenol at 50° C for one hour and then washed in warm water. The enamel could then be removed easily with a cloth. An alternative and simpler treatment was to burn off the enamel by passing a current through the wire. After removal of the enamel the wire was coiled on a $\frac{1}{16}$ in-diameter mandrel and afterwards annealed for two minutes at red heat. Heat treatment at 425° C followed, to bring the temperature-coefficient to the region of zero. The coil was then mounted on to a cylindrical Perspex former and fitted into its case, and after a period in a desiccator it was sealed. No heat treatment was applied after the coil was mounted on its Perspex former since the strains imposed during this operation were considered to be negligible.

The resistance and temperature coefficient of the coil were then measured. The average temperature-coefficient over the range 15–25° C was found to be less than one part in 10^6 per deg C.

The initial resistance was 999·990 ohms on 17th June, 1955. The value has increased by 2 parts in 10^6 during the two months to 16th August, 1955, and by a further 1 part in 10^6 in the following six months to 2nd February, 1956.

(10) 1-OHM STANDARDS

The manganin standards which have proved to have the highest stability have been 1-ohm standards, and a number of 1-ohm standards of Evanohm were constructed for purposes of comparison, although it was realized that this material was not likely to be competitive with manganin for coils of this value. For these standards, wire of 0·048 in diameter was used. The requisite length of wire, less than 100 cm, was bent at its mid-point to form a loop and the double wire coiled round the former. The wire was sufficiently stiff to keep its shape and the former served merely to prevent vibration. It consisted of four grooved Perspex rods secured to a metal cylinder, each rod being provided with a grooved cover of Perspex. The wire was held lightly in the grooves with very little strain. The covers could be taken off and the coil removed for heat treatment without changing the coil shape or applying significant strain. Heat treatment was applied to the shaped coil at 400° C to bring the temperature-coefficient to zero. Five of these standards have been constructed. Their initial values after sealing and their temperature-coefficients are given in Table 6. Heat treatment at 80° C for a few hours was applied to the coils when finally mounted in order to ease any slight strain imposed during this operation.

Table 6

Standard number	Temperature-coefficient over range 15–25° C	Resistance at 20° C on dates given		
		1.8.55	20.9.55	20.2.56
	parts in 10^6 per deg C	ohms	ohms	ohms
L970	+1·0	0·999 931	0·999 933	0·999 933
L971	−0·8	—	1·000 037	1·000 039
L972	+0·3	—	1·000 067	1·000 067
L973	+0·6	—	1·000 019	1·000 023
L974	+0·7	—	0·999 954	0·999 954

(11) OTHER STANDARDS

A number of other resistance standards have been constructed of this material and are in use for general measurements. One of these was described in an earlier paper.[6]

The stability of these resistors has proved satisfactory so far, and there seems little doubt that, provided the wire is not subjected to strain or vibration, the resistance will not change more than a few parts in 10^5 per year. The figures obtained on specially constructed standards indicate the possibility of better stability, but more time will be required before this can be established with certainty.

(12) EFFECT OF COLD WORKING

During the course of the experiments already described, evidence accumulated of the susceptibility of Evanohm to resistance and temperature-coefficient changes caused by strains. A direct experiment was made to determine the effect of coiling wire of 0·004 8 in diameter on to a $\frac{1}{16}$ in-diameter mandrel. It was found that the temperature-coefficient, if near zero, was made more positive by from 4 to 8 parts in 10^6 per deg C. The temperature-coefficient could be brought back to its initial value by heat treatment extending over several days at 200° C, but showed no tendency to do so when a coil was left untouched at room temperature for three months.

It may be observed that the effect of cold working is to move the temperature-coefficient in the opposite direction to that resulting from heat treatment. When, therefore, heat treatment applied to a wire to bring the temperature-coefficient to zero has proved excessive it is tempting to apply cold working or vibration to correct the overshoot. It is likely, however, that the stability of the coil would be adversely affected by such treatment and the alternative method of annealing the wire and repeating the heat treatment is preferable.

In view of the effects of cold working it is undesirable to use Evanohm for resistors which are subject to vibration or rough handling.

(13) CONCLUSIONS

The experiments described show that Evanohm has many desirable qualities as a resistance material. Its stability of resistance, when free from vibration, is good, and for high resistances of fine wire may prove to be superior to manganin. For resistors of moderate value, using wire of substantial section, the stability is also good, but the best figures obtained up to the present are not equal to the best figures for manganin. Better figures may possibly be obtained when the resistors under observation have had a longer settling-down period.

The material is of greatest value for resistors required to operate over a considerable temperature range on account of self-heating, since it is possible to adjust the temperature-coefficient to zero by a simple heat treatment. The change of resistance over the working temperature range resulting from the curvature of the resistance/temperature characteristic is smaller than for constantan and much smaller than for manganin.

The high resistivity, high tensile strength and ductility of the alloy make it very suitable for resistances of high value. For lower values the high resistivity is a disadvantage from the point of view of heat dissipation, but is still an advantage from the point of view of resistance stability since the ratio of surface area to volume is less than that of materials of lower resistivity.

(14) ACKNOWLEDGMENTS

The work described has been carried out as part of the research programme of the National Physical Laboratory, and the paper is published by permission of the Director of the Laboratory. The author desires to acknowledge the assistance rendered in the experimental work by Mr. J. J. Hill and Mr. A. P. Miller.

(15) REFERENCES

(1) PETERSON, C.: 'Alloys for Precision Resistors', Proceedings of a Symposium on Precision Electrical Measurements held at the National Physical Laboratory (H.M. Stationery Office, 1955).
(2) 'New Alloy has Improved Electrical Resistance Properties', *Materials and Methods*, 1948, **28**, p. 62.
(3) WILLIAMS, G. G.: 'Silver Soldering Unit', *Australian Journal of Instrument Technology*, 1947, **3**, p. 110.
(4) RAYNER, E. H.: 'The Effect of Design on the Stability of Manganin Resistances', *Journal of Scientific Instruments*, 1935, **12**, p. 294.
(5) BARBER, C. R., GRIDLEY, A., and HALL, J. A.: 'A Design for Standard Resistance Coils', *Journal of Scientific Instruments*, 1952, **29**, p. 65.
(6) ARNOLD, A. H. M.: 'Alternating-Current-Instrument Testing Equipment', *Proceedings I.E.E.*, Paper No. 1532 M, July, 1953 (**101**, Part II, p. 121).

DISCUSSION ON

'AN EXTENDED ANALYSIS OF ECHO DISTORTION IN THE F.M. TRANSMISSION OF FREQUENCY DIVISION MULTIPLEX'*

Mr. L. Lewin (*communicated*): Messrs. Medhurst and Small state that an echo of amplitude equal to the signal cannot give rise to intermodulation noise, a brief demonstration being given in their Appendix 10.2. This result seems rather surprising, since for small echo amplitudes the distortion certainly increases with echo strength. It also runs counter to one's intuition—admittedly not an infallible guide—of the effects of selective fading on the signal. Since, for equal echo and signal the resulting amplitude modulation is 100%, the conventional limiter cannot deal with the signal in this case. It is therefore interesting to consider in detail the more practical case in which the echo is just a little different in amplitude from the signal.

Accordingly, we take for the combined signal and echo the expression

$$S = \cos(\omega_c t + \phi) + r\cos(\omega_c t + \psi) \quad . \quad . \quad (A)$$

where $\phi = \phi(t)$ is the phase angle representing the frequency-modulated signal and $\psi = \phi(t - \tau) - \omega_c \tau$ is the phase angle of the delayed echo. Eqn. (A) can be put in the form

$$S = A\cos(\omega_c t + \theta) \quad . \quad . \quad . \quad . \quad . \quad (B)$$

where $A^2 = 1 + 2r\cos(\phi - \psi) + r^2$

and $\tan\theta = (\sin\phi + r\sin\psi)/(\cos\phi + r\cos\psi)$

The recovered signal is the instantaneous frequency and is given by $s = d\theta/dt$. From (B),

$$s = \frac{\dot{\phi} + r^2\dot{\psi} + r(\dot{\phi} + \dot{\psi})\cos(\phi - \psi)}{1 + 2r\cos(\phi - \psi) + r^2} \quad . \quad . \quad (C)$$

If we define two angles $\xi = \frac{1}{2}(\phi + \psi)$ and $\eta = \frac{1}{2}(\phi - \psi)$ then (C) can be put in the form

$$s = \xi + \frac{(1 - r^2)\dot{\eta}}{(1 - r)^2 + 4r\cos^2\eta} = \xi + \Delta, \text{ say} \quad . \quad (D)$$

On letting $r \to 1$ we appear to get $s = \xi$, which is the authors' result. However, on closer inspection, we see that the additional term Δ becomes large when η is close to $(n + \frac{1}{2})\pi = \eta_n$, say. In the neighbourhood of η_n, Δ can be written

$$\Delta \simeq \frac{(1 - r^2)\dot{\eta}}{(1 - r)^2 + 4r(\eta - \eta_n)^2} \quad . \quad . \quad . \quad . \quad (E)$$

* MEDHURST, R. G., and SMALL, G. F.: Paper No. 2006 R, March, 1956 (see 103 B, p. 190).

If we integrate Δ through the value η_n we get

$$\int \Delta\, dt \simeq (1-r^2)\int_{\eta_{n-}}^{\eta_{n+}} \frac{d\eta}{(1-r)^2+4r(\eta-\eta_n)^2}$$

$$= \frac{(1-r^2)}{2(1-r)r^{1/2}} \tan^{-1}\left[\frac{2r^{1/2}(\eta-\eta_n)}{1-r}\right]_{\eta_{n-}}^{\eta_{n+}}$$

$$\to \pi \text{ as } r \to 1$$

Thus Δ behaves like the Dirac delta-function, and (D) can be written

$$s = \xi + \pi \sum_n \delta(\eta - \eta_n) \quad . \quad . \quad . \quad . \quad \text{(F)}$$

There is thus a sharp burst of interference every time the angular difference $\phi - \psi$ goes through the value $(2n+1)\pi$, and this must surely be interpreted as intermodulation noise.

The energy contained in these bursts is obtained by integrating the square of the delta function. If we approximate to $\pi\delta(x)$ by

$$\operatorname{Lt}_{\varepsilon\to 0} \varepsilon/(\varepsilon^2+x^2), \text{ then it is easy to see that } \int_{-\infty}^{\infty} [\pi\delta(x)]^2 dx = \pi/2\varepsilon.$$

Thus the energy contained in the interference bursts is very large.

How do these conclusions compare with the authors' other formulae for noise? Their eqn. (11) for the noise in the top channel of a frequency-division multiplex signal contains the expression

$$\mathrm{F}(r) = \frac{r(1-r^2)\sin(\omega_c\tau)}{[1+2r\cos(\omega_c\tau)+r^2]^2} \quad . \quad . \quad . \quad \text{(G)}$$

where $$\cos(\omega_c\tau) = \frac{1+r^2-\sqrt{(1+r^2)^2+32r^2}}{4r} \quad . \quad . \quad \text{(H)}$$

Although this appears to go to zero when $r \to 1$, the denominator actually vanishes too by virtue of eqn. (H). It is therefore necessary to proceed carefully to the limiting value. If we put $r^2 = 1 - \delta$ and retain up to second powers in δ, then it is found that $\mathrm{F}(r) \sim 3\sqrt{(3)}/\delta^2$, which becomes infinite as $\delta \to 0$, i.e. as $r \to 1$.

Thus both methods of analysis lead to the conclusion that the intermodulation noise should be very large when the echo and signal have nearly the same amplitude. It would be interesting to know whether this conclusion is supported by any direct experimental evidence.

Messrs. R. G. Medhurst and **G. F. Small** (*in reply*): We are grateful to Mr. Lewin for drawing attention to the analytical difficulties involved in dealing with the distorting effect of an equal-amplitude echo. It certainly appears that considerable care is needed in going to the limit.

To clear up a minor point first, it is not justifiable to make r go to unity in our expression (11) as Mr. Lewin does in his last two paragraphs. This expression is only valid under the same conditions as those applying to the selective-fading theory of Albersheim and Schafer (our Reference 1). In particular, it requires that $\tau/(1-r)$, where τ is the delay time, should be small. Thus r cannot be allowed to go to unity for a non-zero τ. This is perhaps fortunate, since the infinite distortion/signal ratios derived in Mr. Lewin's last but one paragraph would seem to be as contrary to one's physical intuition as the result of our Section 10.2.

It is, however, undoubtedly true, as Mr. Lewin points out, that in the case of an equal-amplitude echo considerable intermodulation distortion is generated when a certain function of the modulating conditions and the echo delay time is able to exceed a limiting value. For single-tone modulation it turns out that by taking $r < 1$ and going to the limit one can derive fairly simple expressions for the harmonic distortion. These expressions are tedious to write out since their forms depend on the precise ranges of various parameters. As an example, we give the case of the echo in phase quadrature $[\omega_c\tau = (2n+\frac{1}{2})\pi]$. We require a variable A given by

$$A = 2\frac{\omega_D}{\omega_a}\sin(\tfrac{1}{2}\omega_a\tau) \quad . \quad . \quad . \quad . \quad . \quad \text{(J)}$$

where ω_D is the angular frequency deviation and ω_a is the angular modulating frequency. When A is less than $\pi/2$, an equal-amplitude echo generates no harmonic distortion. When $\pi/2 < A < 3\pi/2$, the harmonics in the frequency modulation are given by

$$2\omega_a\left\{\sum_{n=1}^{\infty}\sin\left(2n\sin^{-1}\frac{\pi}{2A}\right)\sin\left[2n(\omega_a t-\tfrac{1}{2}\omega_a\tau)\right]\right.$$

$$\left.-\sum_{n=2}^{\infty}\cos\left[(2n-1)\sin^{-1}\frac{\pi}{2A}\right]\cos\left[(2n-1)(\omega_a t-\tfrac{1}{2}\omega_a\tau)\right]\right\} \quad \text{(K)}$$

When A exceeds $3\pi/2$, additional terms have to be added to the amplitude of each harmonic.

Consideration of the form of this expression clarifies the analysis in our Appendix 10.2, which, in turn, illuminates the mechanism of distortion generation by an equal-amplitude echo. It can be readily seen that the phase modulation distortion obtained by integrating expression (K) with respect to time, consists of a set of rectangular pulses of height π. This is precisely what is implied by the second expression of our Appendix 10.2. In this expression, the 'amplitude modulation' periodically changes sign whenever, in fact, $\omega_c\tau + \phi_{Mt} - \phi_{M(t-\tau)} = (2n-1)\pi$. These changes of sign have to be accounted for by abrupt phase changes of magnitude π. Thus, from the point of view of an ideal limiter and demodulator, the expression should be rewritten as

$$2\left|\cos\tfrac{1}{2}[\omega_c\tau + \phi_{Mt} - \phi_{M(t-\tau)}]\right|\cos\left\{\omega_c t - \tfrac{1}{2}\omega_c\tau\right.$$

$$\left. + \tfrac{1}{2}[\phi_{Mt} + \phi_{M(t-\tau)}] + \mathrm{F}(t)\right\} \quad . \quad \text{(L)}$$

where $\mathrm{F}(t)$ is a pulse function whose Fourier expansion, for the set of conditions assumed in the third paragraph, is given by the time integral of expression (K).

It should, perhaps, be made clear that these limiting considerations in no way affect the results in the main body of the paper where conditions are far removed from the limiting case.

621–526

Paper No. 1895 M
Aug. 1955

AN ON-OFF SERVO MECHANISM WITH PREDICTED CHANGE-OVER

By J. F. COALES, O.B.E., M.A., Member, and A. R. M. NOTON, B.Sc., Graduate.

(*The paper was first received* 11*th February, and in revised form* 14*th May*, 1955. *It was published in August*, 1955, *and was read before the* Measurement and Control Section, 14*th February*, 1956.)

SUMMARY

For a relay-controlled servo mechanism represented by an *n*th-order differential equation it is shown that one change-over and one only, at a unique time, is necessary to bring error and error rate to zero in the least possible time. Previous workers have used the sign of a non-linear function to control change-over, but such methods are suitable only for simple second-order systems with step inputs. To satisfy the need of a technique capable of wider application, e.g. random inputs, prediction of switching by means of a high-speed repetitive analogue computer has been demonstrated with a model experiment. Such a scheme has been shown to be practicable; its use is not limited to simple systems. In the specific example of a 10 kW control system the responses are compared of the on-off control using predicted change-over to those obtained (by simulation) of an orthodox servo mechanism, a linear but saturating servo. The optimized on-off control is always better for all amplitudes of step, ramp and parabolic function inputs, and on the average the same performance would be obtained with both systems using only about half the torque with the on-off control.

(1) INTRODUCTION

When engineers first invented automatic controls it was natural that they should use a simple switch to make a correction whenever the quantity to be controlled was different from the desired value. It was the simplest and most obvious method and was perfectly satisfactory when only a crude control was required. These controls were descriptively referred to as "bang-bang" controls in the United States, "schwarz-weiss" (black-white) in Germany, and as "on-off" controls in this country.

As greater accuracy of control was required, the limitations of simple on-off controls became apparent, e.g. serious overshoot with step inputs and continuous hunting under conditions of no mean error between input and output. Because they are non-linear in the extreme, the general analysis of on-off controls was difficult and exact analysis laborious. However, by going to the increased complication of proportional controls, with derivative and integral terms added, not only could very much improved performances be obtained, but the systems could also be readily analysed and so designed to meet specific requirements.

All components in a control system become non-linear at some level, owing to saturation, and so, to ensure linear operation for ease in analysis, they must all be generously designed; this results in increased size and often in increased complication. During the past few years it has become recognized that, by the appropriate use of non-linear components, control systems can almost certainly be made cheaper and smaller and, in many cases, simpler and more reliable. Unfortunately, the theory of non-linear control systems is exceedingly difficult; thus it has been only slightly explored, and no general methods of analysis or synthesis are available. It is therefore impossible to decide what non-linear components are most likely to be useful. However, on account of their simplicity, on-off controls are immediately attractive and, as a result, a great deal has already been done on the theory.[1,2] They are also attractive for any system in which a large output power has to be controlled. In such a case the most costly part of a control system will always be the output motor, and the cost will be roughly proportional to the peak output power; for this reason the system will usually be designed so that non-linear operation, due to saturation, will first set in as a result of curvature of the torque/input characteristic of the output motor.

If the system is designed for linear operation the maximum torque of the motor will be used only for large values of the error (input minus output), and so for small errors the load will accelerate slowly and will not catch up with the input as quickly as it might. The speed with which the output catches up the input can, of course, be increased by increasing the loop gain of the system, provided it does not become unstable. But this has the effect of reducing the linear region (proportional band), and therefore saturation and non-linear operation will set in for smaller values of the error—such values as are met in normal operation of the system. This usually results in an undesirable increase in the overshoot.

In this connection the on-off controller is effectively an amplifier of infinite gain saturating for infinitesimal values of the error, and so the output motor will always be operated at full torque, either in one direction or the other. Thus, in the case of a position control in which both input and output are at rest, if the input is suddenly moved to a new position (a step function) it is easily seen that the output will be brought into line in the shortest possible time if the full torque of the motor is used to accelerate the load for a certain time and then, in the reverse direction, to decelerate it, so as to bring it to rest exactly at the right position. The problem is, of course, to make the change-over at the correct instant of time.

(2) THE PROBLEM OF OPTIMUM SWITCHING

(2.1) Simple Examples in Second-Order Systems

(2.1.1) Pure Inertia Load.

Consider a constant torque applied in either direction to a pure inertia load, i.e. a frictionless motor and load with ideal torque reversal [Fig. 1(*a*)]. The differential equation of output position can be written in the form

$$b\ddot{X} = L \quad (L = \pm 1) \quad . \quad . \quad . \quad . \quad . \quad (1)$$

If the input is $x(t)$, a function of time but usually abbreviated to x, then the error e at any instant of time is given by

$$e = x - X \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

Eliminating X between eqns. (1) and (2) gives

$$b\ddot{e} + L = b\ddot{x} \quad . \quad . \quad . \quad . \quad . \quad . \quad (3)$$

Let the input as a function of time be given by

$$x(t) = x_0 + \dot{x}_0 t + \tfrac{1}{2}\ddot{x}_0 t^2 \quad . \quad . \quad . \quad . \quad . \quad (4)$$

then eqn. (3) can be rewritten

$$b\dot{e}d\dot{e}/de + L = b\ddot{x}_0 \quad . \quad . \quad . \quad . \quad . \quad (5)$$

Mr. Coales and Mr. Noton are in the Engineering Laboratory, University of Cambridge.

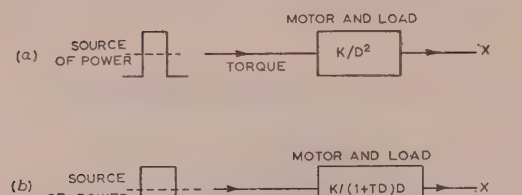


Fig. 1.—Block diagrams for on-off control systems.

(*a*) Second order, pure inertia load.
(*b*) Second order, inertia and viscous friction.
(*c*) Third order, inertia, viscous friction and delay in motor field current ($D = d/dt$).

Because this differential equation is unaltered by replacing e by $e +$ constant, i.e. e is not present explicitly in eqn. (5), all possible solutions, represented as curves in the phase plane of (e, $\dot{e}$), are obtained by shifting any one solution of eqn. (5) parallel to the e-axis. Integration of eqn. (5) gives

$$\tfrac{1}{2}b\dot{e}^2 = (b\ddot{x}_0 - L)e + \text{constant}$$

and the two parabolic curves ($L = \pm 1$) passing through the origin are

$$\tfrac{1}{2}b\dot{e}^2 = (b\ddot{x}_0 \mp 1)e \quad . \quad . \quad . \quad . \quad . \quad . \quad (6)$$

All other trajectories in the phase plane are merely shifts of these two curves parallel to the e-axis.

In the case of step or ramp inputs, $\ddot{x}_0 = 0$ and only the two parabolas

$$\tfrac{1}{2}b\dot{e}^2 = \mp e \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (7)$$

need be considered (for step inputs $\dot{e} = -\dot{X}$). These curves AOA′ and BOB′ passing through the origin are shown in Fig. 2. The portions OA and OB are called the critical trajectories because optimum switching is achieved by reversing the torque when the representative point reaches the curve AOB. Starting from any point in the phase plane P (Fig. 2) the representative

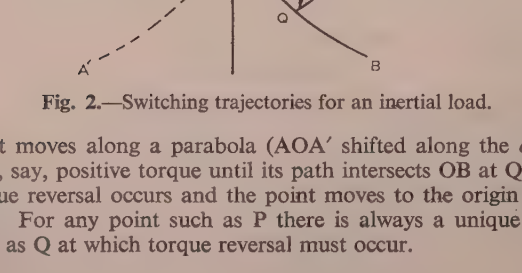


Fig. 2.—Switching trajectories for an inertial load.

point moves along a parabola (AOA′ shifted along the e-axis) with, say, positive torque until its path intersects OB at Q; then torque reversal occurs and the point moves to the origin along QO. For any point such as P there is always a unique point such as Q at which torque reversal must occur.

It is seen from eqn. (7) that optimum switching is achieved by reversing the torque on the sign of the function

$$(\tfrac{1}{2}b\dot{e}^2 \pm e) \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (8)$$

where $-$ applies for $\dot{e} < 0$ and $+$ for $\dot{e} > 0$. This is the switching function to make the best use of the available torque for step or ramp inputs.

The velocity-squared feedback system described by Uttley and Hammond[3] works on this principle, but $\dot{X}^2$ is used instead of $\dot{e}^2$ [eqn. (6)] so that the switching is correct only for step inputs. A similar approach is the S.E.R.M.E. system.[4]

By using a switching function based on eqn. (4) the method could be extended to constant-acceleration inputs. The critical trajectories in the phase plane are still parabolae, but are dependent on the input acceleration.

(2.1.2) Inertia and Viscous Friction.

Consider now a more general second-order system when viscous friction is present [Fig. 1(*b*)]. The differential equation of output position is now

$$b_2\ddot{X} + b_1\dot{X} = L \quad (L = \pm 1) \quad . \quad . \quad . \quad . \quad (9)$$

As before, by substituting $X = x - e$ [eqn. (2)] and restricting the input to the form of eqn. (4), the following equation can be obtained:

$$b_2\ddot{e} + b_1\dot{e} + L = b_2\dddot{x}_0 + b_1(\ddot{x}_0 + \dddot{x}_0 t) \quad . \quad (10)$$

If step inputs only are considered eqn. (10) can be written

$$b_2\dot{e}d\dot{e}/de + b_1\dot{e} + L = 0 \quad . \quad . \quad . \quad . \quad (11)$$

and again, because e is not present explicitly, all possible solutions are obtained by shifting the curves in the (e, $\dot{e}$)-plane parallel to the e-axis. Eqn. (11) is integrated to give

$$\dot{e} - \frac{L}{b_1}\log\left(\frac{b_1\dot{e}}{L} + 1\right) = \text{constant} - \frac{b_1}{b_2}e \quad . \quad . \quad (12)$$

and the two curves passing through the origin are

$$\dot{e} \mp \frac{1}{b_1}\log(1 \pm b_1\dot{e}) + \frac{b_1}{b_2}e = 0 \quad . \quad . \quad . \quad (13)$$

Two such curves are shown in Fig. 3 as AOA′ and BOB′; AOB is the critical switching trajectory.

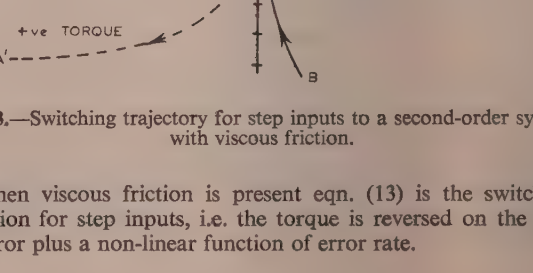


Fig. 3.—Switching trajectory for step inputs to a second-order system with viscous friction.

When viscous friction is present eqn. (13) is the switching function for step inputs, i.e. the torque is reversed on the sign of error plus a non-linear function of error rate.

For constant-velocity inputs the term $b_1\dot{x}_0$ must be retained in the integration of eqn. (10); this leads to the switching trajectory

$$\left.\begin{aligned} \dot{e} - \frac{1 - b_1\dot{x}_0}{b_1}\log\left(1 + \frac{b_1\dot{e}}{1 - b_1\dot{x}_0}\right) + \frac{b_1}{b_2}e = 0,\ \dot{e} > 0 \\ \dot{e} + \frac{1 + b_1\dot{x}_0}{b_1}\log\left(1 - \frac{b_1\dot{e}}{1 + b_1\dot{x}_0}\right) + \frac{b_1}{b_2}e = 0,\ \dot{e} < 0 \end{aligned}\right\} . \quad (14)$$

The non-linear function now involves both error rate and input rate of change.

Several workers[5–8] have demonstrated the transcendental switching function [eqn. (13)] for step inputs to such a second-order system. The non-linear function can be simulated by arrangements of biased diodes, but for ramp inputs [eqn. (14)] the non-linear function depends on the two variables $\dot{e}$ and $\dot{x}_0$, and such a well-known technique is not directly applicable. No work has so far been published on this case.

(2.2) An *n*th-Order System

Consider now a motor and load subject to a torque which is not generally constant. An example of such a case is illustrated in Fig. 1(*c*). Although the voltage across the motor field coils is $\pm$ a constant, the current through the coils is the result of operating on a step function with $1/(1 + TD)$. The applied torque is proportional to that current for a constant armature current. The resulting differential equation is of the third order. It is evident that this can be extended to the general system, which can be represented by a differential equation of the form

$$b_n d^nX/dt^n + \ldots + b_1 dX/dt = L \quad (L = \pm 1) \quad . \quad (15)$$

where the coefficients b_i are usually constants; the following argument is not, however, restricted to that case.

Eliminating X by substituting $X = x - e$ as before gives

$$\sum_{i=1}^{n} b_i d^ie/dt^i + L = \sum_{i=1}^{n} b_i d^ix/dt^i \quad . \quad . \quad . \quad (16)$$

Suppose that the input is prescribed as a function of future time, so that the right-hand side can be written down as a function of time independent of any torque reversal. The problem is to find where a number (P) of torque reversals must occur in order to bring error and error rate to zero in the least possible time.

The differential equation can be solved for the $(P + 1)$ intervals when the applied torque has the same sign. Thus for the first interval, by eliminating t between the e- and $\dot{e}$-equations, the equation of the trajectory in the phase plane of e, $\dot{e}$ can be written in the form

$$h_1(e, \dot{e}, L_1) = 0$$

Similarly for the other intervals the trajectory equations

$$h_2(e, \dot{e}, L_2) = 0 \ldots h_n(e, \dot{e}, L_n) = 0 \quad . \quad . \quad . \quad (17)$$

$$L = \pm 1, L_2 = \mp 1, \text{ etc.}$$

The values of e and $\dot{e}$ are to be calculated at each switch point. For the P switch points there are then $2P$ variables, namely e_1, $\dot{e}_1$, e_2, $\dot{e}_2$, e_3, etc.

$$\text{Variables} - \text{number of equations} = 2P - (P + 1) = a \quad . \quad (18)$$

Assuming the equations are independent there are three cases: (1) $a < 0$, no solution; (2) $a > 0$, no unique solution; and (3) $a = 0$, a unique solution. For the case $a = 0$, $P = 1$, i.e. provided that the initial sign of the torque is correct, for an *n*th-order system one torque reversal only is necessary to bring the representative point in the phase plane (e, $\dot{e}$) to the origin. Once the input is defined exactly as a function of time the solution for the point at which torque reversal must occur is unique and remains unchanged. If, however, the switch point is calculated on the basis of, say, a constant acceleration input, whereas in fact the acceleration changes, so that a higher derivative exists, then an error in switching will result. It is found in specific examples that, if more than one torque reversal is used [corresponding to $a > 0$ in eqn. (18)], the time taken to reach the origin is increased. This is discussed in the Appendix.

Since the work described here started, Bogner and Kazda[9] have published a paper showing that to bring the error and its $(n - 1)$ derivatives to zero in the least possible time $(n - 1)$ reversals are necessary for an *n*th-order differential equation. The above working is a modification of their analysis when following only in position and velocity are required; this is the problem of practical importance.

(2.3) Optimum Switching to Random Inputs

A random input, which in all physical systems must have a finite bandwidth, can always in theory be defined as a function of future time by means of the present values of all its derivatives with respect to time. A formal Taylor expansion can be written down, but owing to the presence of unwanted noise in any system it is usually quite impossible to measure the higher derivatives. It will be assumed that the first and possibly the second can be used. The best calculation of the input as a function of future time is then

$$x(t) = x_0 + \dot{x}_0 t + \tfrac{1}{2}\ddot{x}_0 t^2 \quad . \quad . \quad . \quad . \quad (19)$$

(This is not the "best" in the statistical sense, but such a modification is the subject of proposed future work.) The use of such an input prediction is tantamount to assuming that a random input is composed of a sequence of step functions of acceleration. Because eqn. (19) is an approximation, the calculated point in the (e, $\dot{e}$)-plane at which torque reversal should occur will change slightly as the representative point moves towards the reversal point. Unless the acceleration happens to be constant and eqn. (19) is correct, reversal will not occur at exactly the right point. If the input prediction is the best possible one the switch point is the best estimate that can be made in the circumstances.

It should be noted that, in calculating the switch point from the differential equation in e and x, the derivatives of the input appear because of the corresponding coefficients in the differential equation of motor and load. As most terms have no term depending on output position, the calculation for step inputs when $\dot{x} = \ddot{x} = \ldots = 0$ involves only error and its derivatives. Previous work in this field has consisted almost entirely in developing on-off servo mechanisms with optimum switching only to step inputs of position with corresponding simplification in the switching criteria. For varying inputs, since it is necessary to take into account $\dot{x}$ and if possible $\ddot{x}$, and because viscous friction usually is present, the switching functions will involve derivatives of error and input. The practical realization of switching functions as such then becomes a very formidable problem (and to the knowledge of the authors no work has been published on this important case). In the simple case when only error and its derivatives are required to form the switching function, e.g. eqn. (13), the necessary non-linear operations on $\dot{e}$ are quite practicable. In general, to deal with ramp inputs it is much more difficult [see eqn. (14)], and there is no unique switching trajectory in the phase plane of (e, $\dot{e}$). The switching functions for higher-order systems or for systems with Coulomb friction are still more complicated, whilst higher-order systems require higher derivatives of error (Section 3.3). The general conclusion is that the method of reversing torque on the sign of a switching function is practicable only in the most simple cases.

An alternative approach to the problem of optimum switching, the fast analogue computer technique, has therefore been developed, and its use for the control of a motor and load to varying inputs is described here. The relevant parameters of the second-order system considered represent inertia, viscous and Coulomb friction, but load variations could be introduced, and it is possible to extend the method to higher-order systems.

(3) GENERAL DESCRIPTION OF THE TECHNIQUE

(3.1) Use of a Computer or Simulator

It has been shown that for any on-off system at any instant one switch-over only is necessary to make the trajectory of the representative point in the phase plane (e, $\dot{e}$) pass through the origin of that plane. The co-ordinates of the unique switch-point are required in order to achieve optimum switching. Alternatively, if T_1 is the time ahead when switching must occur, this quantity can be used to control switch-over, for actual switching should take place as T_1 becomes zero. To compute T_1 almost continuously a self-adjusting loop system is used.

This loop system consists of a repetitive computer or simulator working on a very fast time-scale. Starting with the correct initial values this computer gives error and error rate as functions of future time for some time ahead on the fast time-scale. The circuits are then reset and the process repeated, and so on, a large number of computing sweeps being carried out per second. Every computing sweep uses the instantaneous values of the quantity T_1, the switch-over time, written as τ_1 on the fast time-scale.

The computed waveforms of error and error rate are obtained as the difference between input and output waveforms. There is no fundamental difficulty in using, say, an analogue computer to deduce future values of output if the torque, motor and load are defined by an nth-order differential equation and the applied torque is known as a function of time; once $(n-1)$ initial conditions are given the future values of the output are prescribed.

As regards the input the solution is by no means so easy, and, in fact, for a truly random input containing frequencies up to f, the accuracy of any estimate of future values falls off rapidly for times greater than $1/f$ ahead (to an order of magnitude). If, however, a control system is to follow the input, it is presumably designed to respond to the frequency f, and so it would be necessary to predict a time ahead only to the order of $1/f$. For such a prediction in practice eqn. (19) discussed in Section 2.3 has been used.

(3.2) The Computed Switch-Over Time

Consider a typical waveform of error and error rate, on the fast-computer time-scale referred to in Section 3.1, presented as a phase-plane response (drawn in Fig. 4 for a second-order system). The representative point starts at P (present position) and deceleration occurs until the torque is reversed at Q, time τ_1 ahead. The subsequent path does not pass through the origin; torque reversal should have occurred earlier at the point Q′. The value of τ_1 (fast time-scale) or T_1 (actual time-scale) used for this computation of $e(\tau)$ and $\dot{e}(\tau)$ was then incorrect. Let τ_0 be the time when $\dot{e} = 0$ (after switch-over). As seen from Fig. 4 the value of e at time $\tau = \tau_0$ is negative, thus $e(\tau_0) < 0$ implies that τ_1 should be reduced. By using the value $e(\tau_0)$ to change the level τ_1 in the right direction, the next computed sweep of $e(\tau)$ and $\dot{e}(\tau)$ will have a more accurate value of τ_1 and so on until $e(\tau_0) = 0$. The value of τ_1 will then have adjusted itself to make the calculated phase-plane trajectory pass through the origin, i.e. τ_1 is the required switch-over time. Because the period of a computing sweep is very much less than the time-constants of the system, the value of τ_1 is for practical purposes always the optimum switching time. Actual switching occurs when τ_1 becomes zero, and after switching the sequence is repeated.

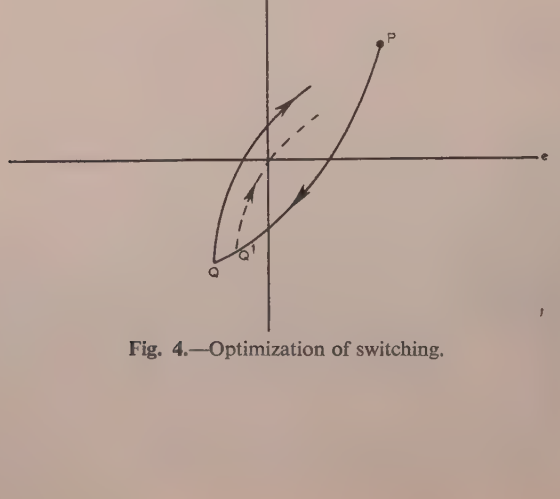


Fig. 4.—Optimization of switching.

The switching rules are not quite so simple as indicated and it is necessary to take into account two factors:

(*a*) In a computed sweep $\dot{e}$ may never be zero, or it may be zero more than once.

(*b*) The change required in τ_1 produced by $e(\tau_0)$ depends in sign on the initial state of the power relay.

A little careful consideration will show that correct switching can be made to depend on the following two rules:

(i) After change-over at τ_1 in the computing sweep e is observed when $\dot{e} = 0$. If after change-over $\dot{e}$ is never zero, e is observed at the end of sweep.

(ii) If the power relay is at $+A$, the value $e < 0$ implies change earlier and $e > 0$ implies change later. If the relay is $-A$, the value $e > 0$ implies change earlier and $e < 0$ implies change later.

(3.3) A System using Optimum Switching

The essential parts of the system as a whole are shown as a block diagram in Fig. 5.

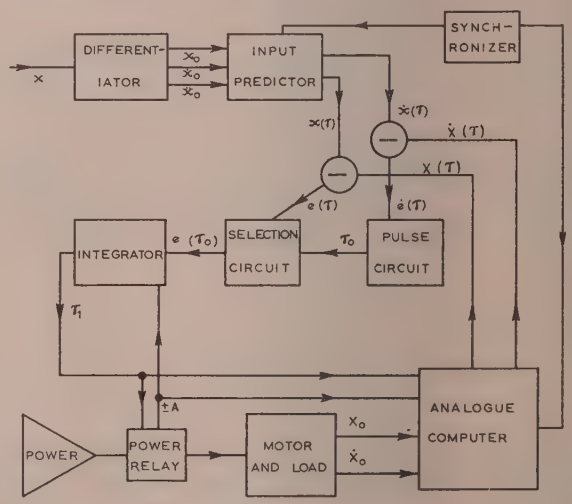


Fig. 5.—Block diagram for a whole system.

The present position of the output X and output velocity $\dot{X}$ are fed continuously to the analogue computer. For higher-order systems, higher derivatives of X are necessary to set the computer, but they need not be measured as such. Take the case of a system which is of the third order due to delay in building up the motor field current [Fig. 1(*c*)]. The simulator block diagram would be the same as that shown in Fig. 1(*c*). To set this

computer the voltages required correspond to X, $\dot{X}$ and the torque. The torque is proportional to the field current which can easily be observed in practice.

The calculated change-over time τ_1 is also fed as a voltage to the computer. Starting with the relevant initial conditions the analogue computer gives a solution to the differential equation of motor and load on the fast time-scale. Torque reversal occurs in this computer at time τ_1; as explained in Section 3.2, this can initially be arbitrary. Waveforms $X(\tau)$ and $\dot{X}(\tau)$ are thus obtained on the fast time-scale and the process is repeated many times a second.

From the input x present values of velocity and acceleration (if possible) are obtained by differentiation. From these are formed waveforms on the fast time-scale of approximate future values of input position and velocity according to

$$\dot{x}(\tau) = \dot{x}_0 + \ddot{x}_0\tau$$
$$x(\tau) = x_0 + \dot{x}_0\tau + \tfrac{1}{2}\ddot{x}_0\tau^2$$

These waveforms are repeated many times a second and synchronized with the output waveforms.

By subtraction $e(\tau)$ and $\dot{e}(\tau)$ are obtained. $\dot{e}(\tau)$ is fed into a pulse circuit which generates a sharp pulse at time τ_0 according to rule (i) of Section 3.2. If the computing frequency is N, there will be N pulses per second.

The pulse serves to select the value of $e(\tau)$ at $\tau = \tau_0$ or at the end of the computing period if $\dot{e}(\tau)$ is never zero in that period, cf. rule (i) of Section 3.2. This pulse $e(\tau_0)$, usually after amplification, becomes the input to an integrator, the d.c. output of which gives the voltage level τ_1. As described in Section 3.2, τ_1 is effectively the output of a self-adjusting loop system, a subsidiary servo system in fact. The loop gain is variable, however, because it depends on the $(e, \dot{e})$-waveforms, and it is a pulsed servo because information is only received in pulses N times a second.

As stated above, when $\tau_1 = 0$ (in fact, when τ_1 passes a critical voltage) the power relay is switched and torque reversal takes place.

(4) EXPERIMENTAL DETAILS OF THE PARTICULAR SYSTEM

A control system has been developed according to the design outlined in Section 3.3. For convenience in analysing the results the motor and load were arranged to have long time-constants, namely 5 sec. The computing frequency was 50 c/s, so that the switching that is necessary in computing circuits at that frequency could be accomplished by high-speed relays. The work was essentially a model experiment; a comparatively small motor and load were used, whereas if this method were applied it would, of course, be to a large high-power control system where the saving in torque could justify the electronic complications introduced.

(4.1) Motor and Load

A Velodyne type 74 was used[10] fitted with a brass flywheel to increase the moment of inertia and consequently the time-constant of the system. The armature shaft rotation was geared down 133 : 1 by means of one worm drive to rotate a special wire-wound potentiometer. The voltage across this potentiometer was a linear function of the angle of rotation and was taken as the output position X_0.

With the connections to the motor as shown in Fig. 6 let the torque $= KI_f i$, where i is the current through the armature A. Let the back e.m.f. induced in the armature coil be $(c/K)I_f\dot{X}$, then

$$(I - i)R = (c/K)I_f\dot{X} + ir$$

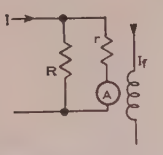


Fig. 6.—Introduction of viscous damping.

Eliminating i,

$$\text{Torque} = KI_f[IR - (c/K)I_f\dot{X}]/(r + R) = J\ddot{X} + F(\dot{X})$$

so that the differential equation of motor and load can be written

$$J\ddot{X} + F(\dot{X}) + cI_f^2\dot{X}/(r + R) = KRII_f/(r + R) \quad . \quad (20)$$

where J is the moment of inertia, c and K are constants and $F(\dot{X})$ is the Coulomb friction term, a term depending only on the sign of the velocity $\dot{X}$ and changing discontinuously at $\dot{X} = 0$. It is seen from eqn. (20) that provided I_f^2 is constant the term in $\dot{X}$ is equivalent to viscous friction. The currents I and I_f (Fig. 6) were made almost independent of motor speed by feeding from the d.c. mains through large resistances. Torque reversal was effected by reversing the direction of the field current I_f, so that I_f^2 was unchanged. The "viscous friction" was varied by means of resistance R and the maximum speed by current I.

Because of the inductance of the field coils the field current cannot, of course, be reversed instantaneously; there is an exponential time-delay which makes eqn. (20) a third-order differential equation. This delay was of the order of 3 millisec in practice. Since the viscous-friction time-constant was 5 sec the 3 millisec delay was insignificant.

Rewriting eqn. (20) as

$$J\ddot{X} + F(\dot{X}) + \lambda\dot{X} = L \quad . \quad . \quad . \quad . \quad . \quad (21)$$

L being the applied torque, one solution for $\dot{X}$ can be written

$$\left.\begin{aligned} \dot{X} &= V(1 - \varepsilon^{-t/T}), \quad \text{for} \quad \dot{X} > 0 \\ \dot{X} &= V'(1 - \varepsilon^{-t/T}), \quad \text{for} \quad \dot{X} < 0 \end{aligned}\right\} \quad . \quad . \quad (22)$$

where
$$\left.\begin{aligned} V &= (L - F)/\lambda, \text{ the maximum speed} \\ V' &= (L + F)/\lambda \\ T &= J/\lambda \end{aligned}\right\} \quad . \quad . \quad (23)$$

If T' is the time to decelerate from top speed to rest, it is easily shown from eqns. (22) and (23) that

$$F/L = 2\varepsilon^{-T'/T} - 1 \quad . \quad . \quad . \quad . \quad . \quad (24)$$

This relation is useful for observing indirectly the ratio of Coulomb friction to applied torque, F/L.

By means of a pen-recorder measuring velocity $\dot{X}$ from the tacho-generator, the actual performance of the motor and load was measured and the differential equation (21) verified. The parameters V, T and T' were found to have these values:

$V = 910$ r.p.m. of the armature shaft or one rotation of the potentiometer shaft in 8·8 sec.
$T = 5{\cdot}0$ sec.
$T' = 2{\cdot}8$ sec.

The other parameters of the system were approximately:

Total moment of inertia $= 29 \times 10^3$ gm-cm^2.
Applied torque = 660 gm-cm.
Coulomb torque = 95 gm-cm.

The field current of 24 mA was switched by a relay.

(4.2) The Analogue Computer

The circuit for deriving $X(\tau)$ and $\dot{X}(\tau)$ is shown as a block diagram in Fig. 7. $\pm L(\tau_1)$ denotes the applied torque which

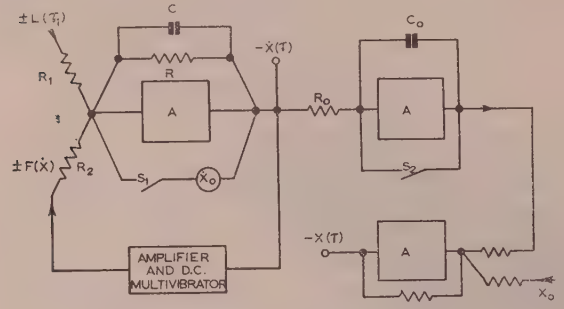


Fig. 7.—Simulation of motor and load.

reverses at τ_1 and $\pm F(\dot{X})$ the Coulomb frictional torque. The block marked A denotes a high-gain d.c. amplifier. The circuit equation for the first part can be written

$$\pm L/R_1 \pm F/R_2 = -(1 + RCD)\dot{X}/R \quad . \quad . \quad (25)$$

where $D \equiv d/dt$. The differential equation for $\dot{X}(\tau)$ is then

$$RC\ddot{X} \pm (R/R_2)F(\dot{X}) + \dot{X} = \pm (R/R_1)L(\tau_1) \quad . \quad (26)$$

which by correct choice of components can be made equivalent on the fast time-scale to the actual equation of motor and load, eqns. (20) and (21).

The duration of each computation was 0·01 sec, which corresponded to 10 sec actual time. Successive computing sweeps were separated by equal periods for resetting. During a resetting period the high-speed relays S_1 and S_2 are closed, so that when a computing sweep starts, by opening the relays, $\dot{X}(\tau)$ starts at X_0 and $X(\tau)$ at zero, X_0 the present position being added to $X(\tau)$ in the next stage. $\dot{X}_0$ is the voltage developed by the tacho-generator attached to the motor shaft, and its impedance is negligible compared to R_1 or R_2.

The operation from $\dot{X}(\tau)$ to $X(\tau)$ is pure integration. Provided that AR_0C_0 is very much greater than the computing period, the second stage is pure integration, followed in the third stage by addition of X_0. Thus

$$X(\tau) = (1/R_0C_0)\int_0^\tau \dot{X}(\tau)d\tau + X_0 \quad . \quad . \quad . \quad (27)$$

The applied torque was simulated by the output voltage of a bistable multivibrator. At the beginning of each computing sweep the state of the multivibrator was set to correspond with the state of the power relay. After a time τ_1 in each sweep the multivibrator state was changed (the analogue of torque reversal) by means of a trigger pulse generated when the voltage output of a conventional Miller time-base coincided with the voltage level τ_1. The time-base was itself triggered from the common synchronizing source.

The voltage corresponding to the torque of Coulomb friction was also derived from a direct-coupled multivibrator, but its state depended only on the sign of the computed velocity $\dot{X}(\tau)$.

(4.3) The Input Predictor

The input predictor was built entirely with standard plug-in amplifiers and is shown in its basic form schematically in Fig. 8. The input velocity is derived from the input by a feedback differentiator, and similarly the input acceleration from the input

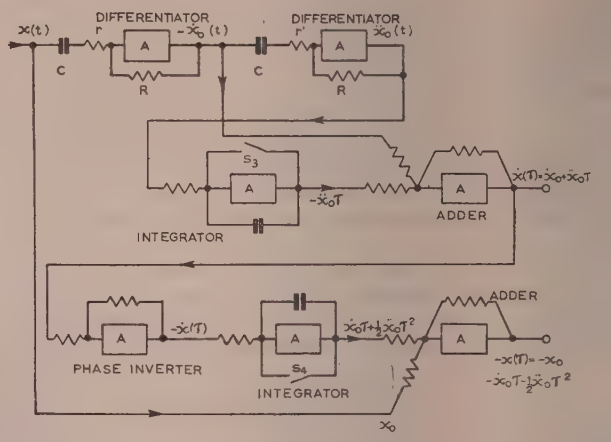


Fig. 8.—Input predictor.

velocity. The differentiation was purposely made imperfect by the resistors r and r′, i.e. the first circuit gives differentiation with an exponential delay-time rC and the second circuit has a delay $r'C$. This is equivalent to filtering the input and was found essential in order to lessen the effect of transients on power supplies, etc. The delay for velocity was 1/20 sec and for acceleration 1/5 sec. Special precautions were also taken to decouple these differentiator circuits from the other circuits. Although these differentiators were satisfactory for the ideal test inputs described, modification will probably be necessary for subsequent work.

For slowly varying inputs containing frequencies up to, say, 0·1 c/s, the voltages x_0, $\dot{x}_0$ and $\ddot{x}_0$ are virtually d.c. levels for the period of one computing sweep. The integral $\int_0^\tau \ddot{x}_0 d\tau$ can then be written $\ddot{x}_0\tau$ and similarly for other integrations. By correct choice of time-constants the functions

$$\dot{x}(\tau) = \dot{x}_0 + \ddot{x}_0\tau$$
$$x(\tau) = x_0 + \dot{x}_0\tau + \tfrac{1}{2}\ddot{x}_0\tau^2$$

were generated by processes of addition and integration as shown in Fig. 8 for the fast time-scale. Again the high-speed relays S_3 and S_4 must be synchronized with a common source and consequently with S_1 and S_2 of the analogue computer.

(4.4) Pulse Circuit

Having formed $\dot{e}(\tau)$ by addition of $\dot{x}(\tau)$ and $-\dot{X}(\tau)$ the waveform was amplified and applied to a multivibrator comparator circuit[11] which gave a sharp pulse whenever $\dot{e}(\tau)$ passed a voltage level corresponding to $\dot{e}(\tau) = 0$.

To set up rule (i) of Section 3.2 a combination of four multivibrators was used as indicated in Fig. 9. Multivibrator 1 generates a rectangular pulse about 0·2 millisec long when $\dot{e} = 0$, but the circuit is paralysed by a voltage from the analogue computer until after computed change-over at τ_1. Multivibrator 2, triggered from the main synchronizing source, gives a sharp pulse just before the end of the computing sweep. This pulse triggers multivibrator 3 to generate a rectangle pulse 0·2 millisec long provided that multivibrator 1 has not fired. If multivibrator 1 has fired then multivibrator 3 is paralysed to the end of that sweep by a waveform from multivibrator 4. In this way a rectangular pulse 1/50th of the computing period in length is generated at time τ_0 according to rule (i) of Section 3.2.

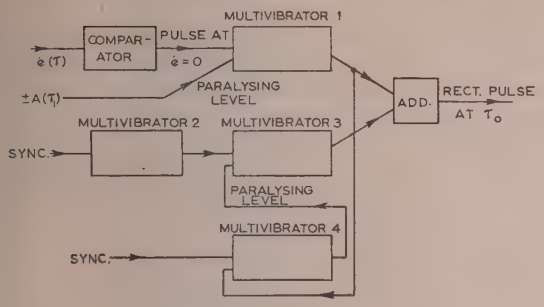


Fig. 9.—Pulse circuits for τ_0.

(4.5) Selection Circuit

The waveform $-e(\tau)$ was obtained by adding $-x(\tau)$ and $X(\tau)$, and the value of $e(\tau)$ at $\tau = \tau_0$ was achieved by a conventional selection circuit[12] shown in Fig. 10. To prevent overloading of

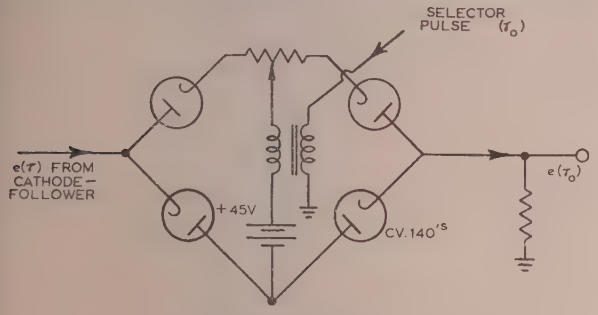


Fig. 10.—Selection circuit.

this selection circuit the magnitude of $e(\tau)$ was limited to ± 40 volts, i.e. clamping diodes came into action outside this range of voltage. The non-linearity introduced makes little difference to the performance of the system for changing τ_1.

(4.6) Integrator

In order to satisfy rule (ii) of Section 3.2 it was necessary to invert the phase of the pulse $e(\tau_0)$ for one position of the power relay. Phase inversion was achieved by a feedback amplifier of unit gain switched into circuit when necessary by a Post Office relay.

After amplification the pulse $\pm e(\tau_0)$ was fed into a simple one-valve integrator, the anode voltage of which was τ_1. When this voltage passed a critical value a polarized relay closed to generate a trigger pulse for a bistable multivibrator. The power relay was operated by the anode current of one of the triodes of that multivibrator, so that the torque was reversed every time τ_1 passed the critical value $\tau_1 = 0$ from positive to negative.

In order to return the voltage below the critical value a direct-coupled multivibrator came into action, injecting a voltage at the grid of the integrator until the anode voltage was well below the critical value once more.

(5) EXPERIMENTAL RESULTS

(5.1) Circuit Drifts, etc.

The stability of the circuits was such that the equipment could be used satisfactorily 10 min after the heaters were switched on.

The stabilized h.t. supply voltages were regulated within $\pm 0{\cdot}7\%$.

Signal voltages were kept sufficiently high throughout to limit errors due to valve drift voltages to about $\pm 2\%$. Through the use of negative feedback, errors due to non-linearities in the amplifiers were negligible. Computing circuits were lined up by adjusting components to an accuracy of $\pm 1\%$.

The instant at which the high-speed polarized relays opened was adjusted by the current through a separate bias coil, the relays being driven by a 50 c/s sine wave. Depending on the voltage and current being switched, some of the relay contacts needed cleaning and respacing after less than one hundred hours' use. In general, therefore, it would seem desirable to use electronic switching.

(5.2) Performance

With no input the output hunted in a limit cycle about the zero-error position. The response to a transient input of position and/or velocity consists of a period using $\pm$torque, a period using $\mp$torque followed by hunting about $e = 0$, $\dot{e} = 0$.

Since the system was designed to respond in the best possible manner to step functions of acceleration, zero acceleration being a special case, observations were first made of the responses to step functions of position, velocity and acceleration. The inputs were generated by one or two integrators. The responses were observed with a pen recorder the time-constant of which was about one-quarter of a second.

The results for several inputs are reproduced in Figs. 11–15.

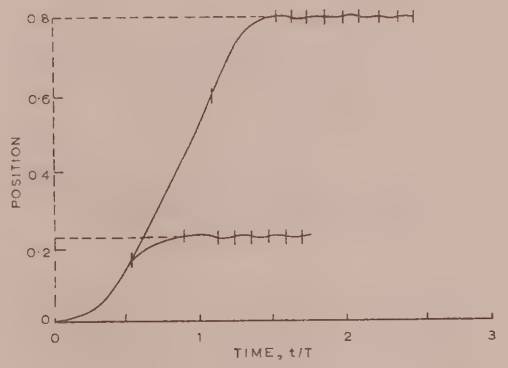


Fig. 11.—Responses to step inputs.

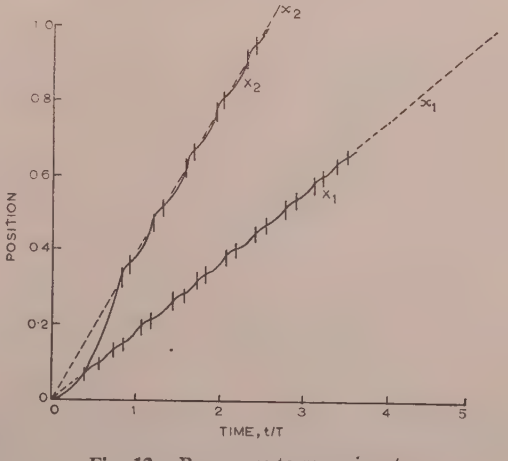


Fig. 12.—Responses to ramp inputs.
$\dot{x}_1$ is 0·16 V and $\dot{x}_2$ is 0·34 V.

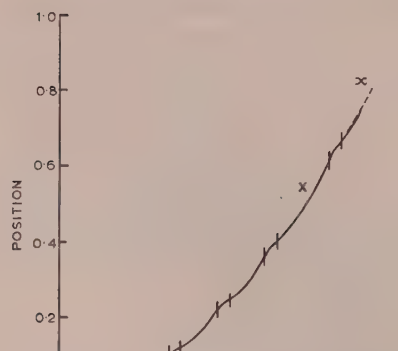


Fig. 13.—Response to parabolic input.
$\dot{x}$ at time T is $0{\cdot}18V$.

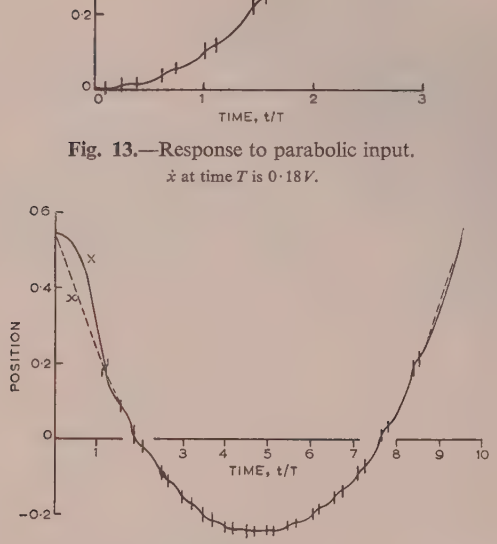


Fig. 14.—Response to parabolic input (initial error of velocity).
x is $-0{\cdot}36V$ initially and $0{\cdot}36V$ at time $9{\cdot}5T$.

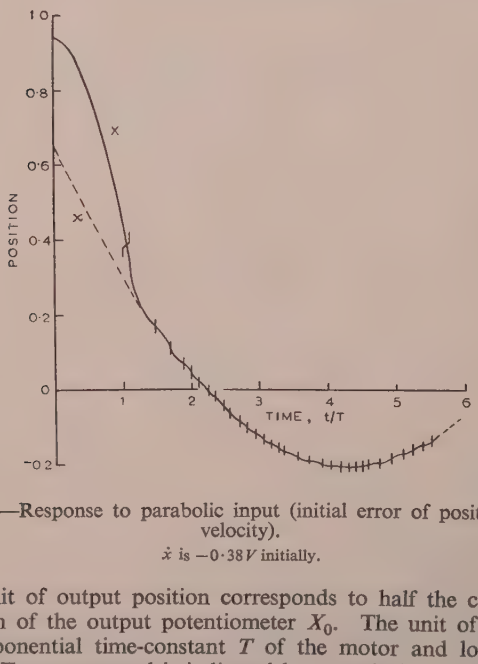


Fig. 15.—Response to parabolic input (initial error of position and velocity).
$\dot{x}$ is $-0{\cdot}38V$ initially.

The unit of output position corresponds to half the complete rotation of the output potentiometer X_0. The unit of time is the exponential time-constant T of the motor and load, viz. 5 sec. Torque reversal is indicated by a vertical line across the response curve and the value of V quoted on the figures is the maximum speed of the motor.

With an ideal system the hunting frequency would become infinite, but in practice inevitable delays in switching are present. Backlash in the gears and delays in the relays were present, but by far the most important delay was that involved in computing the change-over time. This delay can be reduced by increasing the gain in the loop which computes τ_1, but because that loop forms a pulsed servo system the gain cannot be increased indefinitely;[13] however, the higher the computing frequency the higher the maximum permissible gain.

Two factors must, however, be taken into account.

(*a*) The drift variations in the critical value $\tau_1 = 0$ (at which torque reversal occurs) due to h.t. supply variations and contact wear in the high-speed relays, corresponded to a fluctuating error in switching time of about 0·08 sec.

(*b*) Torque reversal before the correct time is undesirable; for, by consideration of the result of always switching a given short interval of time early or late, it can be shown that the time to reach the origin of the phase plane is always less in the latter case. Slightly late switching gives rise to overshoot, but when switching takes place too early the representative point approaches the origin of the phase plane $(e, \dot{e})$ along a zig-zag path, as illustrated in Fig. 16 (compare under-damped and over-damped responses

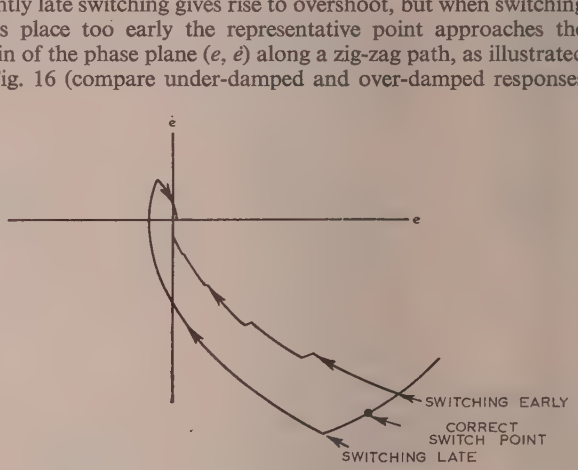


Fig. 16.—Effect of early and late switching.

of a linear proportional servo mechanism). Because a random error in change-over time of 0·08 sec was present, to ensure a slight overall delay in switching it was necessary to have a delay in the system greater than 0·08 sec. It was convenient to have a delay of 0·15 sec in computing, and this gave a limit to the gain of that loop. If that loop gain were increased, it would only be necessary to introduce a delay elsewhere.

It is shown in Section 8.2 that a delay T_{delay} in reversing the torque after the ideal instant (neglecting friction which is permissible for small oscillations) gives rise to hunting oscillations of period T_{osc} where

$$T_{osc}/T_{delay} = 13{\cdot}7$$

e.g. with a switching delay of 0·15 sec the period of oscillation is 2·1 sec. This was borne out in practice, but the hunting period varied somewhat due to the fluctuations in T_{delay} discussed above. This was in fact the only appreciable effect on the performance of the system due to drift variations.

(5.3) Comparison with an Orthodox Servo Mechanism

It is of considerable interest to take stock of the position to see just what has been gained in making this very close approximation to the optimum servo mechanism. Could the electronic complications be justified? How do the responses compare with those which would be achieved by an orthodox servo mechanism,

e. a proportional linear system possibly with saturation. The omparison can be expected to depend on amplitude, because at east one system is highly non-linear.

Let the model motor and load be scaled up, but with the same ime-constants, to a control system using a maximum power of, ay, 10 kW. Assume the time-constant of the field coils to be ·2 sec; the delay associated in computing change-over time ould be reduced so that this 0·2 sec delay is the only appreciable elay in the system. The hunting period would be about $2\frac{1}{2}$ sec, nd the responses would be as given in Figs. 11–15.

Now consider the control of the same motor and load (i.e. ame moment of inertia, same viscous friction, but neglecting Coulomb friction for simplicity) using a proportional servo nechanism with the torque limited to the same value used in he above on-off servo mechanism. The two systems will then ave the same maximum velocity, and the torque characteristic s then linear with saturation. For such powers it would, how-ver, be necessary to use some form of rotary amplifier (with nevitable time-delays). Suppose a metadyne be used to give a ower amplification from 50 watts, derived electronically, to 0 kW; it would have a transfer function of the form

$$K/(1 + T_f p)(1 + T_a p)$$

where p is Laplace's operator, T_f is the time-constant of the exciting field, and T_a the combined time-constant of the metadyne armature and the field coil of the motor. Typical values[14] for the power gain of 200 would be $T_f = T_a = 0{\cdot}2$ sec.

Provided that the non-linear saturation effect does not become serious it is desirable to use as high a loop gain as possible. To obtain stability it will, however, be necessary to introduce a

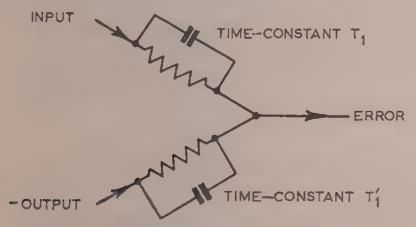


Fig. 17.—Phase advance network.

phase advance network (Fig. 17). Assume that this has the ideal form of transfer function

$$(1 + T_1 p)$$

i.e. the unwanted delay of such networks has been made negligible. This is always possible, for ideal inputs, by means of a feedback amplifier circuit. If ω is the natural frequency of the system and T the viscous friction time-constant as before (see Fig. 17), by making

$$T_1 = T_1' + \frac{1}{\omega^2 T}$$

the servo mechanism will have no persistent error to constant-velocity inputs.[15]

The maximum loop gain can now be fixed by considering the stability of the whole system. When in the linear region the differential equation is

$$\left(\frac{T_f^2}{\omega^2}p^4 + \frac{2T_f}{\omega^2}p^3 + \frac{1}{\omega^2}p^2 + T_1 p + 1\right)X = (1 + T_1 p)x \quad . \quad (28)$$

For a damping ratio of 2/3, $T_1\omega = 3/2$, giving the characteristic equation

$$\frac{T_f^2}{\omega^2}p^4 + \frac{2T_f}{\omega^2}p^3 + \frac{1}{\omega^2}p^2 + \frac{3}{2\omega}p + 1 = 0 \quad . \quad . \quad (29)$$

The Routh–Hurwitz condition for stability leads to

$$T_f\omega < 12/25$$

Taking $T_f\omega = 0{\cdot}2$ (to obtain a well-damped response), $\omega = 1\,\text{sec}^{-1}$, $T_1 = 1{\cdot}5$ sec.

The response of this system with torque limitation to several inputs has been calculated and is shown in Figs. 18–20. The

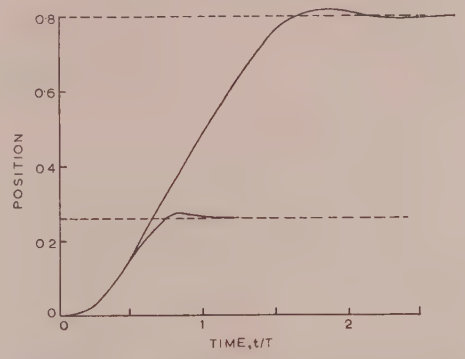


Fig. 18.—Responses of orthodox servo mechanism.

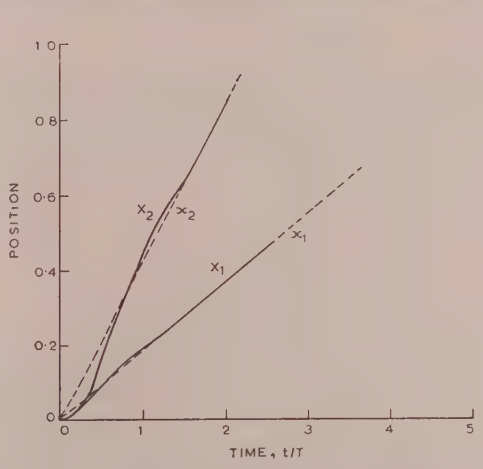


Fig. 19.—Responses of orthodox servo mechanism.
x_1 is 0·16 *V* and x_2 is 0·38 *V*.

responses were, in fact, obtained by simulating the system with several feedback-amplifier circuits, diode rectifiers being used to give the torque limitation. The inputs were approximately the same as used for the on-off servo mechanism, so that a direct comparison can be made between the two systems.

The corresponding responses for the optimum on-off and orthodox servo mechanism are Figs. 11 and 18, 12 and 19, 15 and 20. The most satisfactory comparison is to plot the time taken for both systems to reduce the error to a given small value against the magnitude of the initial error. These curves have been calculated for step inputs and are shown in Fig. 21. Strictly, the curve for the orthodox servo mechanism is more irregular than shown because of the damped oscillations, but the dotted line is approximately correct. For small inputs the orthodox servo mechanism behaves as a linear system. The optimum on-off servo mechanism is always better than the other, but the big difference occurs for small-amplitude inputs where the comparison is between a linear and a highly non-linear system, e.g. for a step input of 0·05 unit of position the on-off system takes

Fig. 20.—Responses of orthodox servo mechanism.
$\dot{x}$ is $-0{\cdot}35V$ initially.

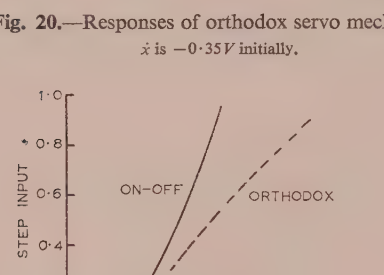


Fig. 21.—Comparison of optimum on-off and orthodox (saturating) servo mechanism for different-amplitude step inputs.

0·4 unit of time, whereas the orthodox system takes 0·7 unit. The same comparison for the other inputs would lead to similar conclusions. The responses shown are typical.

A "rough and ready" conclusion is that, on the average, the on-off servo mechanism brings input and output into line in about two-thirds of the time taken by the orthodox servo mechanism. Because time taken is proportional to the square root of the applied torque, it follows that the torque of the on-off system could be reduced by half to give a similar performance to the orthodox servomechanism. The maximum speed of the former would, however, be halved.

The choice of ω from the stability criterion may seem approximate, but it was verified on the simulator that even if ω could be increased by increasing the loop gain while the system remained stable, the response could not be improved due to the saturation effect. The linear region of the torque characteristic was too narrow, and the response was tending to that obtained with an on-off servo mechanism switching on a line in the phase plane $e + T_1\dot{e} = 0$.

These conclusions must necessarily be rather tentative at this stage, but at least they do give an indication of the advantages of this on-off servo mechanism. The only really satisfactory comparison is to control a system already in existence by this means.

(6) CONCLUSIONS

A general method has been devised for achieving optimu switching with an on-off control system. The practicability c predicting the ideal switching time has been demonstrated wit a model experiment for which responses to step, ramp, an parabolic input functions have been found to compare favourabl with those of an orthodox system.

Using the hypothetical example of a 10 kW control system, th probable saving in torque which could result if the optimu on-off servo mechanism were used has been deduced. The mo powerful the control system considered the more importa becomes this saving in torque, and the complications of th associated electronics can then be justified. For example, a power levels of the order of 100 h.p. a 50% saving in torqu (apart from the elimination of rotary power amplifiers) suggest that this method of on-off control may well have practica applications.

The design of a high-power control system would fall int two stages: (*a*) subject to practical limitations the torque reversa is made as rapid as possible; the delay in reversing the torqu may be simply that due to the inductance of the field coils, bu whatever it is the second stage is (*b*) to simulate the system an the reversing torque function actually present. The best use i then being made of the available torque and mode of reversa It is, of course, quite feasible to limit the rate of build-up c torque in any desired way, since this can be taken into accoun in the simulation of the actual torque function.

In conclusion it may be stressed again that, once the fas analogue computer is made, the additions required to the com puter in order to simulate highly complicated motor, load an torque characteristics are insignificant. Furthermore, there is n need to know theoretical governing equations, since the compute can be made to conform to characteristics obtained from test carried out over the working range of the system. Changes i characteristics of motor and load can easily be taken into accoun during life by changes to the computer.

It is envisaged that future work will include

(*a*) An introduction of dead-zone or torque reversal in a finit time to eliminate the wasteful drain of power in hunting,

(*b*) Simulation of the character of torque reversal likely to b met in practice,

(*c*) The application of statistical methods of input predictio for dealing with random inputs in the presence of noise,

(*d*) Modifications to take into account load variations.

(7) REFERENCES

(1) HAZEN, H. L.: "Theory of Servomechanisms," *Journal of th Franklin Institute*, September, 1934, **218**, p. 279.

(2) WEISS, H. K.: "Analysis of Relay Servomechanisms, *Journal of the Aeronautical Sciences*, July, 1946, **13**, p. 364

(3) HAMMOND, P. H., and UTTLEY, A. M.: "The Stabilizatio of On-off Controlled Servomechanisms," Automatic an Manual Control (Butterworth, 1952), p. 285.

(4) WEST, J. C., DOUCE, J. L., and NAYLOR, R.: "The Effect of the Addition of Some Non-linear Elements on th Transient Performance of a Simple R.P.C. System possess ing Amplifier Saturation," *Proceedings I.E.E.*, Pape No. 1549 S (**101**, Part II, p. 156).

(5) HOPKIN, A. M.: "A Phase-plane Approach to the Com pensation of Saturating Servomechanisms," *Transaction of the American I.E.E.*, 1951, **70**, Part I, p. 631.

(6) SCHWARTZ, J. W.: "Piecewise Linear Servomechanisms, *ibid.*, 1953, **71**, Part II, p. 401.

(7) SILVA, L. M.: "Non-linear Optimisation of Relay Servos, Thesis, University of California, Berkeley, California 1953.

(8) NEISWANDER, R. S., and MACNEAL, R. H.: "Optimisation of Non-linear Control Systems by Means of Non-linear Feedbacks," *Transactions of the American I.E.E.*, 1953, **72**, Part II, p. 262.
(9) BOGNER, I., and KAZDA, L. F.: "An Investigation of the Switching Criteria for Higher-Order Contactor Servomechanisms," *ibid.*, 1954, **73**, Part II, p. 118.
(10) WILLIAMS, F. C., and UTTLEY, A. M.: "The Velodyne," *Journal I.E.E.*, 1946, **93**, Part IIIA, p. 1256.
(11) CHANCE, B., WILLIAMS, F. C., etc.: "Waveforms" (M.I.T., Radiation Laboratory Series), Vol. 19, p. 341.
(12) *Ibid.*, p. 374.
(13) JAMES, H. M., NICHOLS, N. B., and PHILIPS, R. S.: "Theory of Servomechanisms," *ibid.*, p. 231.
(14) TUSTIN, A.: "D.C. Machines for Control Systems," (Spon, 1952), p. 66.
(15) WEST, J. C.: "Textbook of Servomechanisms" (English Universities Press, 1953), p. 68.

(8) APPENDICES

(8.1) Proof that the One Torque Reversal gives Minimum Time

Although this fact may seem obvious it is difficult to find a precise mathematical proof in a general form. The following argument, although semi-qualitative, should serve to resolve any doubts about the truth of the statement.

It has been shown that, to bring the representative point in the phase plane (e, $\dot{e}$) to the origin, only one torque reversal is necessary at a unique point for an nth-order system. Is it possible by using more than one torque reversal to reach the origin in a shorter time? This problem is more conveniently discussed in terms of the phase plane of (X, $\dot{X}$) rather than the error phase plane; for if it can be shown that the time of passing from any point P to any point Q in the phase plane of (X, $\dot{X}$) is shorter using one torque reversal than by any other path using more than one torque reversal, the theorem in the (e, $\dot{e}$)-plane follows.

Take first the second-order system

$$b_2\ddot{X} + b_1\dot{X} = L(t) \quad . \quad . \quad . \quad . \quad . \quad (30)$$

where the torque $L(t) = \pm 1$. The phase-plane trajectories of this system have been discussed in Section 2. Characteristic of a second-order system are two and only two trajectories through any point in the phase plane (X, $\dot{X}$) for the two directions of the torque $L = \pm 1$. Neglecting the cases when the torque has the wrong sign initially, the typical cases that might arise are discussed with reference to Fig. 22. The unique path with one

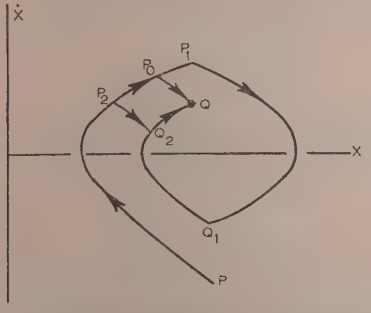


Fig. 22.—Phase-plane trajectories for a second-order system.

torque reversal is PP_0Q, positive torque from P to P_0 and negative torque from P_0 to Q. Two other paths are considered, PP_1Q_1Q and PP_2Q_2Q, each using an extra torque reversal. Clearly as PP_2 is common to all three cases this portion can be neglected. The time taken between any two points (1, 2) is $\int_1^2 dX/\dot{X}$. Comparing the paths P_2P_0Q and P_2Q_2Q, because $\dot{X}$ is always less along the latter path the integral for that path is greater, and so all such paths as P_2Q_2Q are over a longer time-interval than that via P_0Q. To examine paths such as $P_0P_1Q_1Q$ consider the integral of eqn. (30) between the points P_0Q:

$$b_2(\dot{X}_2 - \dot{X}_1) + b_1(X_2 - X_1) = \int_1^2 L(t)dt = (t_2 - t_1)\bar{L}(t) \quad . \quad (31)$$

where $\bar{L}(t)$ is the mean value of the torque in the time interval t_1 to t_2. Eqn. (31) can be rewritten

$$t_2 - t_1 = [b_2(\dot{X}_2 - \dot{X}_1) + b_1(X_2 - X_1)]/\bar{L}(t) \quad . \quad (32)$$

For the case in the diagram $\bar{L}(t) < 0$, but the time of passing from P_0 to Q varies inversely as $|\bar{L}(t)|$, since the numerator of eqn. (32) is prescribed. The torque has constant sign for the path P_0Q, but the path $P_0P_1Q_1Q$ involves reversals and so quite clearly $|\bar{L}(t)|$ must be greatest for the unique path P_0Q (compare Fig. 24). The path involving one reversal is then the path of shortest time for a second-order system.

The higher-order systems of interest arise from delays in applying the torque, e.g. the torque of eqn. (30) becomes for a third-order system

$$L(t)/(1 + TD), \qquad D \equiv d/dt$$

which indicates an exponential delay, of time-constant T, as $L(t)$ takes the values ± 1. In on-off control systems only such values of L are relevant, and so the discussion can be limited to second-order systems provided that the modified form of $L(t)$ is taken into account. Consider Fig. 23, which is the modification of Fig. 22

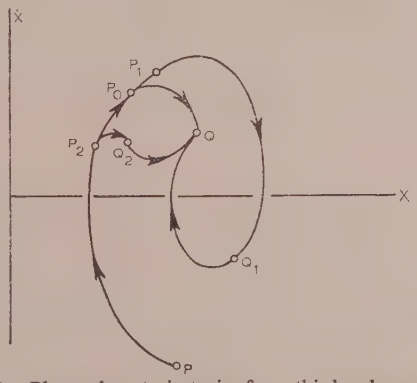


Fig. 23.—Phase-plane trajectories for a third-order system.

for a third-order system. PP_0Q is the path with one torque reversal, and the other two paths involve extra switchings. Because the torque cannot now be reversed instantaneously, "switch" is the preferred term. Characteristic of a third-order or higher-order system through every point of the phase plane there are an infinite number of possible trajectories. It is seen that when switching occurs the slope of paths in the plane is now continuous. For the same reason as before the path P_2Q_2Q is obviously over a longer time interval than the path P_2P_0Q. For the path $P_0P_1Q_1Q$ the same argument can be used again. Eqn. (32) still applies provided the modified form of $L(t)$ is used. The torque variations for the paths P_0Q and $P_0P_1Q_1Q$ are shown in Fig. 25. Again $\bar{L}(t)$ is negative for the case illustrated but clearly $|\bar{L}(t)|$ cannot be greater for the path $P_0P_1Q_1Q$ than for the path P_0Q, and so the latter is the path of shortest time for a third-order system.

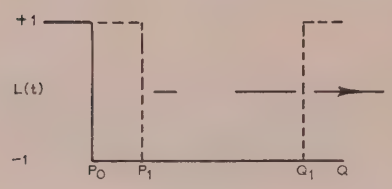


Fig. 24.—Torque variations for a second-order system.

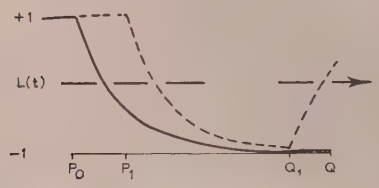


Fig. 25.—Torque variations for a third-order system.

Although the torque variations will be modified, the above argument may be extended to higher-order systems.

(8.2) Effect of Switching Delay

Consider the small hunting oscillations of an on-off servo mechanism in the quiescent state where the torque is not switched at the ideal instant but a short interval of time later. For simplicity Coulomb friction is neglected, and for small oscillations the effect of viscous friction is negligible. It has been shown for a second-order system in Section 2.1 that ideal switching occurs on a unique trajectory composed of two parabolae in the phase plane $(X, \dot{X})$ [compare eqn. (7)]. If time t is taken as the parameter the two parabolae can be represented by

$$X = \pm at^2, \ \dot{X} = \pm 2at \quad . \quad . \quad . \quad . \quad . \quad (33)$$

where hunting is about the point $X = 0$. Other trajectories a merely shifts of these parallel to the X-axis.

Stable oscillation (a limit cycle) is represented by PQRS

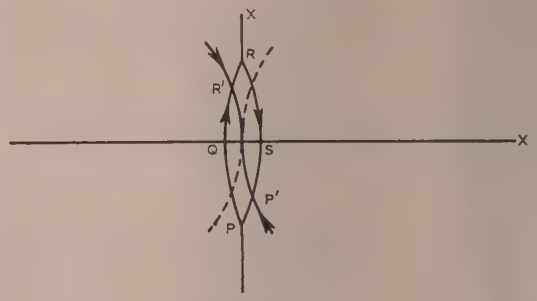


Fig. 26.—Effect of switching delay.
Ideal switching at P′, R′.
Actual switching at P, R.

Fig. 26; the closed curve must, of course, be symmetrical abo the $\dot{X}$-axis. The trajectory PQR is given by

$$X = a(t^2 - t_1^2), \ \dot{X} = 2at \quad . \quad . \quad . \quad . \quad . \quad (3$$

where t_1 is the time at point R ($X = 0$), $t = 0$ being the point ($X = 0$). Now R′ is given by the common point of eqns. (3 and (34) if R′ corresponds to time t_2.

$$X(R') = a(t_2^2 - t_1^2) = -at_2^2$$

Therefore $t_2/t_1 = 1/\sqrt{2}, \quad (t_1 - t_2)/t_1 = (\sqrt{2} - 1)/\sqrt{2}$

The period of oscillation $T_{osc} = 4t_1$ and the switching delay $T_{delay} = t_1 - t_2$.

Therefore $T_{osc}/T_{delay} = 4\sqrt{2}/(\sqrt{2} - 1) = 13{\cdot}7 \quad . \quad (3$

Almost the same ratio is derived by means of describing functio

DISCUSSION BEFORE THE MEASUREMENT AND CONTROL SECTION, 14TH FEBRUARY, 1956

Dr. A. M. Uttley: In this system the authors suggest differentiating the error, and in the past this has been rather a dangerous thing to do. In eqn. (8) the switching function is $b\dot{e}^2 \pm e$. In practice there is noise coming in, and to differentiate the error will greatly magnify that noise on its way to the switching point. To avoid this, Mr. Hammond and I tried transient feedback of $v|v|$ so as to differentiate the output from a motor, which cannot be as lively as a noisy signal. This is crude mathematically, but it worked well for step functions, and I believe it would be possible to apply it to random functions. I should like the authors' views on this.

The authors have suggested that their system is valuable for large power systems, but I wonder whether they are 'underselling' their method and whether its great point is that it produces a better servo mechanism rather than that it saves money. It also may be the case that the switching problem will become progressively more difficult as the powers become larger. I do not know whether it is practical to switch 100 h.p. motors.

In Section 8.1 the authors refer to the interesting fact that the one torque reversal gives minimum time, and say that this seems obvious but that it is difficult to prove. From the first paragraph of the Appendix, I wonder whether the authors are happy that their proof would satisfy a Hardy. Do all the coefficients ha to be positive for this theorem to be true? I remain uncerta whether or not this theorem has been proved.

Dr. J. C. West: I should like to emphasize the point raised the last paragraph of Section 3.1. Consider first a simple line servo-system with equation

$$\left(\frac{p^2}{\omega_0^2} + Tp + 1\right)\theta_0 = \theta_i$$

It is the terms on the left-hand side, due to inertia and visco forces, which prevent the ideal system

$$\theta_0 = \theta_i$$

If, however, we introduce a slight circuit modification, such phase advance, the input to the system can be made to be t sum of the desired quantity θ_i and a term proportional to rate change $Tp\theta_i$. The system equation is

$$\left(\frac{p^2}{\omega_0^2} + Tp + 1\right)\theta_0 = \left(1 + Tp\right)\theta_i$$

and has become one stage nearer perfection. If now an acceler

ion of the desired quantity is measured and also fed into the ystem, the equation might become

$$\left(\frac{p^2}{\omega_0^2} + Tp + 1\right)\theta_0 = \left(\frac{p^2}{\omega_0^2} + Tp + 1\right)\theta_i$$
$$\theta_0 = \theta_i$$

nd perfection would be achieved. There are two ways of egarding this modification of θ_i to form a control signal. The irst is to regard the transfer functions on the right-hand side as eing desirable to help to cancel the undesirable but inevitable ransfer functions on the left-hand side. The second is to regard he modified input signal as some form of prediction, in that, ince velocity and acceleration are taken into account, the signal an represent the future position of the system at a time ahead ot greater than the system response time.

The ideal system is never achieved for many reasons, but nainly because:

(*a*) No system is of second order completely, and therefore higher orders of input derivatives would be required.
(*b*) Perfect differentiation to obtain the required characteristics is not physically realizable—it can only be approached.
(*c*) Noise in the input with high frequencies present will produce large signals in the control circuits and swamp the system.
(*d*) This is the most important point: the torque required to make the system more and more perfect approaches infinity, and all practical systems have a maximum available torque, i.e. they are non-linear to the extent that torque saturation inevitably takes place.

It is this limitation of torque which limits the performance of a system. For an inertia system it is possible to draw a straight line on the closed-loop frequency-response diagram falling at a rate of 12 dB/octave, which represents the ultimate limit of performance for a given maximum torque. The essence of the paper is an elegant method to get nearer to this practical limit. The $V|V|$ and Serme type of systems approach this quite closely, even for sinusoidal and random signals. The authors' method achieves greater perfection. It is therefore necessary to decide in a given application the importance of the extra degree of optimization and to balance this with the extra complications of an additional fast analogue simulator.

Dr. J. P. Corbett: In the Conclusions it is stated that the on/off system with predicted change-over may have advantage over a linear system for large sizes of equipment. So far as the transient response is concerned, the paper has adequately shown this to be the case. However, the introduction of a predicted change-over will inevitably lead to high-frequency oscillation of the system in the steady state. The wear and tear due to this on contacts or other switching devices in large servo systems will be serious. For this reason, I believe that as both systems have advantages, arrangements incorporating both principles should be sought. The ideal can be approached by having a system with a small linear zone and saturation at as low a value as is convenient to give the required torque in the saturated condition.

Consider a system such as is shown in Fig. A(i) having a characteristic as shown by the broken line in Fig. A(ii). The equalizing arrangements for such a system, treated on a linear basis, must be compatible with the need for a certain value of predicted change-over time.

A device producing the predicted change-over effect will have to come before the amplifier, as shown in Fig. A. (I feel that feedback functions can be neglected because, as shown in the paper, they compensate only for given types of input disturbances.)

It has been shown that, to obtain the best transient performance, the error signal requires to be modified in the on/off case so that

$$\beta_1 = K_1\varepsilon + K_2\dot{\varepsilon}^2 + K_3\dot{\varepsilon}^3 \ \ldots$$

where β_1 is a proportion of the new error signal.

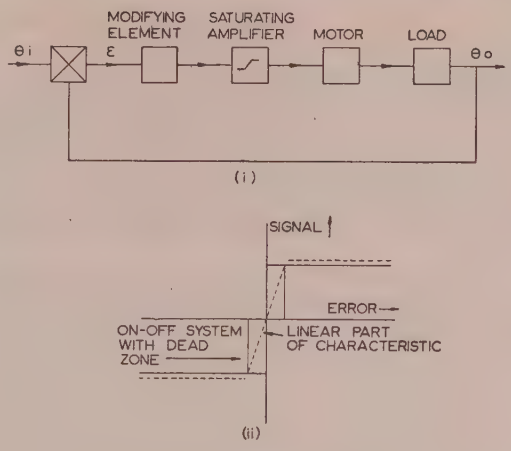


Fig. A.—Quasi-linear servo system.
(i) Block diagram.
(ii) Forward-loop steady-state characteristic.

For stability in the linear region the error requires to be modified so that

$$\beta_2 = K_a\varepsilon + K_b\dot{\varepsilon} + K_c\ddot{\varepsilon} \ \ldots$$

where again β_2 is a proportion of the new error signal.

Thus, in the one case the sum of the modulus of the integer powers requires producing in various proportions, and in the other the sum of the derivatives in various proportions is required.

The best possible system will be obtained by a compromise between these requirements, and I feel this compromise will be arrived at most rapidly with a simulator.

Dr. J. L. Douce: In a conventional servo mechanism stabilized by derivative-of-error damping the differentiating unit will, in general, overload when a step input is applied. This effect may occur for different signal levels, dependent on the sign of the applied step, so that the step response depends on the sign of the applied input. In the paper the input signal is differentiated twice, and the output of the predictor may differ appreciably from its expected value. Does this produce a noticeable departure from the desired response? What modifications do the authors intend in Section 4.3 for use when test inputs other than their ideal are being applied to the system?

It is surprising that emphasis is placed on the improvement effected in the step response of the new system. For a step input the best predictor may well be no predictor at all, since input velocity and acceleration are both zero for all time after the step has been applied.

An improvement of approximately 2 : 1 is quoted for the optimized system for a particular small input step. However, this step is only twelve times the hunting amplitude, giving a final error of 8%, which may not be acceptable in practical applications.

The equivalent conventional servo mechanism appears to have been designed with the gain limited by time-constants ($\simeq 0{\cdot}2$ sec) associated with a metadyne amplifier. With commercially available magnetic amplifiers a gain of 200 can be obtained with a time-constant of 0·02 sec or less, so that such a system would compare much more favourably with the optimized system.

The authors mention the incorporation of field build-up time delays into the simulator, so that the simulator may then be representing an unstable system. Does the present technique fail in these circumstances?

[The authors' reply to the above discussion will be found overleaf.]

THE AUTHORS' REPLY TO THE ABOVE DISCUSSION

Messrs. J. F. Coales and **A. R. M. Noton** (*in reply*): The switching function of eqn. (8) involving error-rate will clearly give rise to the same switching for step inputs as the $v|v|$ system recalled by Dr. Uttley. For constant-velocity inputs, however, the latter system will exhibit a persistent error, cf. the velocity-lag of linear servo-mechanisms stabilized by means of output-velocity feedback. It is for this reason that the velocity of the input should be taken into account, if possible, as well as that of the output.

Drs. Uttley, West and Douce all questioned the possibility of differentiating inputs which, in practice, are so often contaminated with unwanted noise. Since the paper was written, considerable attention has been given to the problem of noisy random inputs. Provided that the statistical properties of signal and noise are known, in order to achieve the best prediction, a systematic treatment of the input is possible by means of Wiener's theory of noise filtering. The resulting predictions make use of as many derivatives as the random input possesses, generally with preliminary filtering of the noise. Perhaps an example is not out of place.

Consider a random input composed of a sequence of random steps of velocity. Such an input is not a stationary random process, but its derivative may be considered as such. Furthermore, if the time interval between the step changes of velocity varies according to a Poisson distribution, it can be shown that the spectral density of the input derivative is of the form

$$\Phi_D(\omega) = \frac{1}{(\omega_0^2 + \omega^2)}$$

Let the unwanted noise added to the random signal-input have a constant-frequency spectrum—so-called 'white noise'—

$$\Phi_{nn}(\omega) = n^2$$

Assuming that signal and noise are uncorrelated, the use of Wiener's theory leads to the best predicting transfer function to operate on the input, namely

$$F(p) = \frac{1}{(1 + Np + np^2)} \left[1 + p/\omega_0 - p/\omega_0\, \varepsilon^{-\omega_0\tau}\,(1 + N\omega_0 + n\omega_0^2)^{-1}\right]$$

where p is Laplace's operator, $N = \sqrt{(2n + n^2\omega_0^2)}$ and the prediction is for a time τ ahead. A special case of interest arises when n is small and $\tau \ll 1/\omega_0$, i.e. prediction is required only for a time ahead much less than the average interval between changes of rate. Then

$$F(p) = \frac{(1 + \tau p)}{(1 + Np + np^2)}$$

Since $F(p)$ operates on the input, $x(t)$ say, the above transfer function can be either produced *en bloc*, or the operation may be regarded as two stages:

(a) Filter with the transfer function $1/(1 + Np + Np^2)$

(b) From this filtered signal x_0', form its derivative $\dot{x}_0'$ and predict according to

$$x(\tau) = x_0' + \dot{x}_0'\tau$$

This particular signal input possesses, of course, only on derivative.

Even without the approximations the operation of predictio may be divided into two such stages; the predicted position (an velocity) is then the best possible in the circumstances, yet th modifications to the input predictor described in the paper ar by no means drastic.

Contrary to the suggestion (a) of Dr. West, the number o derivatives of the input taken into account in the prediction i independent of the order of the motor and load but depends o how many derivatives the random signal input possesses. A indicated above, the presence of unwanted noise will not, c course, permit pure differentiation, and preliminary filterin would be necessary. We agree with him that in many case particularly in the case of small relatively simple motor and loads the $V|V|$ or Serme system would be adequate and that th complications of the repetitive simulator would be unjustified These non-linear feedback networks were derived, however, fo second-order systems, so that an exponential time-lag in reversin the torque (due to the inductance of, say, the armature coils would be troublesome. Such a time lag, which makes th system of the third-order, can, however, be taken into accoun in the method of predicted change-over, and so the method i applicable to higher-order systems.

We appreciate the doubts of Drs. Uttley and Corbett on th persistent switching of high-powered motors; there is clearl scope for research in this direction. In order to lessen th induced e.m.f.'s which give rise to arcing, the rate of torqu reversal could presumably be purposely limited and the mode o torque reversal be imitated in the repetitive simulator. Such modification has been shown to cause only a slight deteriorati of the transient performance.

Dr. Corbett's suggested use of saturating systems certainl avoids the switching problems, but the starting-point of th research described in the paper was to employ relay controls i order to eliminate power amplifiers. The introduction of dea zone is possibly a more attractive way of preventing persisten hunting oscillations under conditions of no input. It is tru however, that there is scope for using saturating systems in dual mode of operation. For large errors, when the torque, sa is saturated, the sudden reversal of torque is determined by non-linear switching function or predicted change-over, and fo small errors the system is operated as a conventional pro portional control.

In answer to Dr. Uttley's query, we are not satisfied that th proof in the Appendix would satisfy a Hardy; unfortunately, proof depending entirely on analytical methods has not yet bee devised.

As Dr. Douce points out, the equivalent conventional serv mechanism of Section 5.3 was designed with the gain limited b time-constants associated with the metadyne amplifier. It was however, verified on a simulator that, even if the unwanted time lag was reduced, the gain could not be increased, owing to th effect of saturation. As a result of saturation, a compromise i setting the gain had to be made for different magnitudes of ste inputs.

In answer to the last question, the incorporation of fiel build-up time delays into the simulator does not lead to a unstable system, since in the ordinary sense the simulator doe not represent a closed-loop system.

;21-52

Paper No. 1877 M
July 1955

THE DUAL-INPUT DESCRIBING FUNCTION AND ITS USE IN THE ANALYSIS OF NON-LINEAR FEEDBACK SYSTEMS

By J. C. WEST, Ph.D., Associate Member, J. L. DOUCE, Ph.D., Graduate, and R. K. LIVESLEY, Ph.D.

The paper was first received 20th September, 1954, in revised form 28th January, and in final form 15th April, 1955. It was published in July, 1955, and was read before the MEASUREMENT AND CONTROL SECTION *13th March, 1956.)*

SUMMARY

A new method is presented for the frequency-response analysis of non-linear elements. This involves evaluating the gain of one of the frequency components in passing through the non-linear element when the input to the element consists of two sinusoidal waves of differing amplitudes, phases and frequencies. A cubic characteristic is analysed fully as an example.

The paper describes the use of this method for analysing four different types of problem arising in feedback systems containing one simple-type amplitude non-linearity. The method can be considered as an extension of the describing-function technique with a correspondingly wider field of application.

(1) INTRODUCTION

In recent years there has been a growing tendency to make use of steady-state frequency-response methods in the analysis of non-linear feedback systems. This has been limited to systems possessing low-pass filtering elements (usually inherent) following the non-linearity in the feedback loop, which effectively remove the harmonics introduced by the non-linearity.[1] In these cases it is possible to assume that the feedback signal is approximately sinusoidal and consequently that the error signal and the input to the non-linear element are also sinusoidal. On this assumption it is possible to define the gain of the non-linear element for a sinusoidal signal as

$$G(a) = \frac{\text{amplitude of fundamental component of output}}{\text{amplitude of input}}$$

which is dependent on the input amplitude a. $-1/G(a)$ is known as the describing function. Techniques have been developed[2,3,4] in which the describing function is used for considering the natural stability of a system when no input is applied, and for determining the steady-state frequency response of the closed-loop non-linear system. In each case it is reasonable to assume a sinusoidal signal being applied to the non-linear element. To widen the field of application of such techniques it becomes necessary to consider a more complex signal fed to the non-linear element, and the first approach is that of a dual-frequency signal

$$v_i = a \cos(\omega t + \phi) + b \cos n\omega t \quad . \quad . \quad . \quad (1)$$

the only restriction on a, b and n being that they must be real. The output from the non-linear element now contains beat-frequency components in addition to harmonics of the individual input frequencies. However, the dual-input describing function can be defined as the ratio of the amplitude of the desired frequency component (required for the analysis) in the output waveform to the amplitude of the component of the same frequency in the input waveform. This becomes a function of the four variables a, b, n and ϕ.

The dual-input describing function is calculated for the case of cubic non-linearity, to illustrate the methods employed. To evaluate the response of a saturation-type non-linear element a digital computer has been employed, and general response curves for this type of element are presented.

Several uses have been found for the dual-input describing function, and the technique provides a quantitative explanation for the phenomena introduced below.

For a lightly-damped system, it has been shown that the "jump effect" may occur;[3,4] i.e. a small variation in the input amplitude or frequency is accompanied by a large discontinuous change in output amplitude. In the range of input for which this jump occurs, the theory of the steady-state response of the system predicts three possible values of the output amplitude. In practice only two of these values are encountered, so that the third solution must be regarded as unstable. Steady-state analysis provides no method of considering the stability of the forced oscillations.

The technique presented enables the response of the closed-loop system to be determined for small sinusoidal signals in the presence of the primary input of fixed frequency and superimposed on the steady-state solution. This enables the incremental Nyquist diagram to be plotted, and consideration of the diagram gives the stability of the system to any small additional disturbance.

It is shown that this response locus may enclose the $(-1, 0)$ point for certain values of error amplitude and frequency. Thus any slight disturbance under these conditions produces a large effect on the output amplitude. It is shown that the range of error amplitude and frequency in which the response of the system is unstable coincides with the region in which the jump effect is predicted.

It is also found that there is a region on the complex plane such that, if the open-loop frequency-response locus of the linear part of the system passes through this region the jump effect can occur. This region is evaluated for two particular cases of non-linearity, and experimental results are presented giving good agreement with theory.

When a sinusoidal signal is applied to a "simple" non-linear element,[8] i.e. one whose output is a single-valued function of the input amplitude, it may be shown that the fundamental component of the output is in phase with the input.[3] If the input consists of two sinusoidal signals whose frequencies are in simple relation, both signals may undergo a change in phase in passing through the non-linear element.

A feedback system under these conditions may be unstable, the frequency of oscillation being simply related to that of the primary input sinusoid. These oscillations are known as harmonics or sub-harmonics,[5] depending on whether the frequency of oscillation is greater or less than the input frequency. Sufficient information is available from the Nyquist diagram of the linear system to determine whether or not these forced oscillations can occur, and the amplitude of any possible oscillation.

A further application of the technique presented in the paper concerns the stability of a conditionally stable system. Experi-

Dr. West and Dr. Douce are in the Electrical Engineering Laboratories, University of Manchester.

Dr. Livesley was formerly in the Computing Machine Laboratory, University of Manchester, and is now in the Engineering Laboratories, University of Cambridge.

ment shows that for small inputs such a system is stable but that for some critical value of input amplitude the system bursts into continuous oscillation, the frequency of oscillation being independent of the input frequency. The theory shows how the gain of the non-linear element for a small signal of one frequency is modified by the presence of a further signal of unrelated frequency. It follows that continuous oscillations are started by a given amplitude of signal applied to the non-linear element, independent of the frequency of the forcing signal. From this, the variation of critical input amplitude with frequency may be determined by known techniques.

Section 5 describes a fourth application of the dual-input describing function in order to estimate some of the inaccuracies involved when the normal describing function is used.

(2) EVALUATION OF THE DUAL-INPUT DESCRIBING FUNCTION

To determine the dual-input describing function it is necessary to find the output waveform of the non-linear element for an input of the form

$$v_i = a \cos(\omega t + \phi) + b \cos n\omega t \quad . \quad . \quad . \quad (1)$$

and to analyse the output to find the components of a particular frequency.

When the non-linear characteristic can be expressed as a simple power series an algebraic analysis is straightforward, and the function is derived as a relatively simple expression involving a, b, n and ϕ.

It is desirable, however, to extend this technique to the analysis of practical feedback systems, where the inherent non-linearity is rarely expressible as a simple power series. An important example of unavoidable non-linearity is given by symmetrical saturation, and the algebraic analysis of this non-linear element is difficult. Hence other methods of deriving the required data are investigated.

A graphical analysis is possible in general, but is too tedious in view of the large number of sets of results required—one thousand for the response for ten values of a, b and ϕ. Experimental methods have been employed in which input and output are passed through tuned filters to isolate the desired components which are then compared in amplitude and phase.

Alternatively, a more accurate and flexible method is to determine the required describing function by means of a digital computer. The Manchester University machine permits the response of the non-linear element to be determined in about 14 sec, and the complete describing function requires about ten hours of computing time.

Having computed the response of the non-linear element, it is possible to use the computer to evaluate the response of the complete closed-loop system, for example, by means of the time-series technique for the linear system.[1,2] This method which considers the waveform as a series of spot values instead of as a fundamental plus third harmonic, is capable of giving results to any desired accuracy. The method also differs from the describing-function technique in that unstable solutions are never produced; the computer behaves in a manner identical to that of the original system. The disadvantage is that the response must be recalculated for every change either in input or in any parameter of the system. It is considered preferable to present general results giving the response of the non-linear element alone and thus permit analysis of any system where symmetrical saturation provides the non-linear element.

(2.1) The Cubic Non-Linearity

Consider an input signal given by eqn. (1) applied to the non-linearity $v_0 = v_i^3$.

The output signal is

$$v_0 = [a \cos(\omega t + \phi) + b \cos n\omega t]^3 \quad . \quad . \quad . \quad ($$

which on expansion becomes

$$\frac{3a}{4}(a^2 + 2b^2) \cos(\omega t + \phi) + \frac{3b}{4}(2a^2 + b^2) \cos n\omega t$$

Table 1

TERMS OF THE DUAL-INPUT DESCRIBING FUNCTION FOR A CUBIC CHARACTERISTIC

Application	Frequency required	Amplitude restrictions	Frequency conditions	Approximations made	Relevant term
Stability of forced oscillations of a non-linear feedback system ..	$n\omega$	$b \ll a$	$n \neq 1$	Terms in b^2 and b^3 neglected	$\frac{3}{2}a^2b \cos n\omega t$
			$n = 1$	Ditto	$\frac{3}{4}a^2b[(2 \cos \omega t + \cos(\omega t + 2\phi)]$
Synchronized oscillations and sub-harmonic resonance	ω	None	$n \neq \frac{1}{3}, 1$ or 3	None	$\frac{3a}{4}(a^2 + 2b^2) \cos(\omega t + \phi)$
			$n = 3$	None	$\frac{3a}{4}(a^2 + 2b^2) \cos(\omega t + \phi) + ab \cos(\omega t - 2\phi)$
			$n = \frac{1}{3}$	None	$\frac{3a}{4}(a^2 + 2b^2) \cos(\omega t + \phi) + \frac{b^3}{4} \cos \omega t$
Self-maintained oscillations of conditionally stable systems	$n\omega$	$b \ll a$	$n \neq 1$	Terms in b^2 and b^3 neglected	$\frac{3}{2}a^2b \cos n\omega t$
Effect of third harmonic on natural stability of non-linear feedback system	ω	$b < a$	$n = 3$	$\frac{b^2}{a^2}$ and $\frac{b^3}{a^3}$ neglected	$\frac{3a^2}{4}[(a \cos \omega t + \phi) + b \cos(\omega t - 2\phi)]$

$$+\frac{a^3}{4}\cos 3(\omega t+\phi)+\frac{b^3}{4}\cos 3n\omega t$$

$$+\frac{3a^2b}{4}\{\cos[(n+2)\omega t+2\phi]+\cos[(n-2)\omega t-2\phi]\}$$

$$+\frac{3ab^2}{4}\{\cos[(2n+1)\omega t+\phi]+\cos[(2n-1)\omega t-\phi]\} \quad (3)$$

In particular when n is any value except $\frac{1}{3}$, 1 or 3 the component of frequency $\omega/2\pi$ is

$$\frac{3a}{4}(a^2+2b^2)\cos(\omega t+\phi) \quad . \quad . \quad . \quad (4)$$

The input component of this frequency is $a\cos(\omega t+\phi)$ and hence the apparent gain is

$$\tfrac{3}{4}(a^2+2b^2) \quad . \quad . \quad . \quad . \quad . \quad . \quad (5)$$

and there is no phase change. When b is zero, i.e. for a single-frequency input, the gain reduces to $3a^2/4$ and is the normal describing function for a cubic characteristic.

Different components are required for different applications, and in some cases approximations may be made by restrictions on amplitudes and frequencies. For the four cases dealt with in the paper, Table 1 has been constructed to show the relevant conditions and the required components.

(2.2) Symmetrical Saturation

Consider the non-linear characteristic representing symmetrical saturation, defined by

$$v_0 = v_i \qquad -1 \leqslant v_i \leqslant 1$$
$$v_0 = -1 \qquad v_i < -1$$
$$v_0 = \ \ 1 \qquad v_i > \ \ 1$$

+4	+1				
+0·3239	+3·1521	+0·0281	+2·4324	− 0·1124	
+0·3235	+3·1527	+0·1000	+2·4359	− 0·4475	
+0·3217	+3·1548	+0·1656	+2·4512	− 0·9323	
+0·3168	+3·1607	+0·2182	+2·4942	− 2·0729	
+0·2986	+3·1838	+0·2288	+2·6652	+28·2942	
+0·2134	+3·3057	+0·1103	+3·8711	+ 0·9417	
+0·1950	+3·3292	−0·0671	+4·2667	− 0·5174	
+0·2842	+3·2035	−0·2142	+2·8180	+ 5·0850	
+0·3138	+3·1645	−0·2277	+2·5209	+ 2·7737	
+0·3206	+3·1561	−0·1818	+2·4610	+ 1·1033	
+0·3231	+3·1531	−0·1159	+2·4392	+ 0·5516	
+0·3241	+3·1520	−0·0446	+2·4313	+ 0·1956	
FRACTION OF HARMONIC IN OUTPUT →	$\frac{1}{\text{GAIN}}$ OF FUNDAMENTAL COMPONENT →	TANGENT OF PHASE CHANGE OF FUNDAMENTAL COMPONENT →	$\frac{1}{\text{GAIN}}$ OF HARMONIC COMPONENT →	TANGENT OF PHASE CHANGE OF HARMONIC COMPONENT →	INCREASING VALUES OF ϕ

Fig. 1.—Data as printed by computer for $a = 4$, $b = 1$.

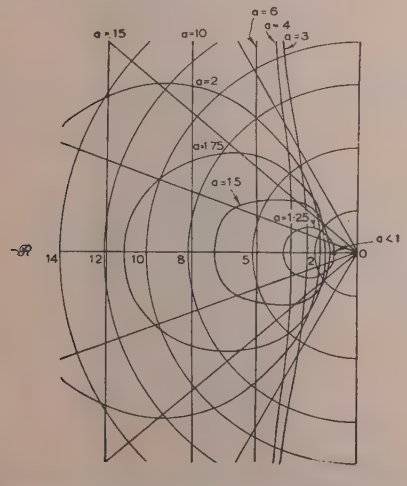
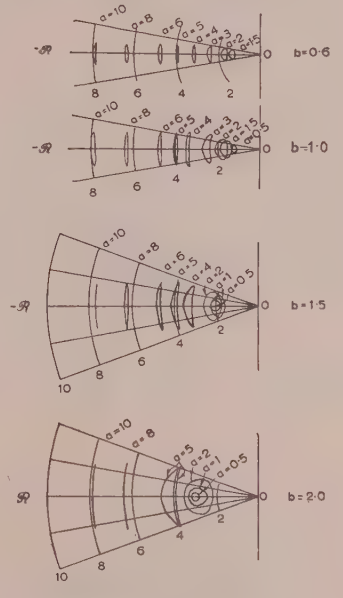


Fig. 2.—Incremental describing-function loci for saturation.



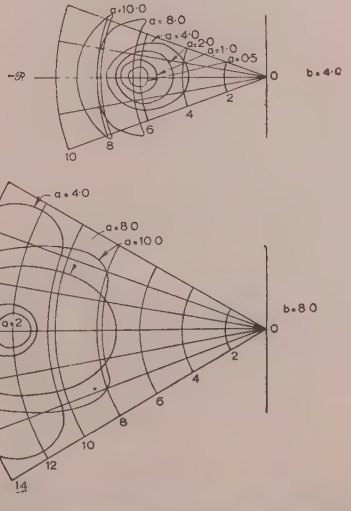


Fig. 3.—Low-frequency describing-function loci.

The general input to the non-linear element is taken, as before, to be of the form

$$v_i = a \cos(\omega t + \phi) + b \cos n\omega t \quad . \quad . \quad . \quad (1)$$

where n is a small integer.

For $n = 1$ and $b \ll a$, the describing function for the component involving b is the incremental describing function.

Since a sinusoidal input to the non-linear element produces no second harmonic at the output, the next analysis is for $n = 3$. In this case the describing functions for the components of frequencies $\omega/2\pi$ and $3\omega/2\pi$ are both required.

Cases of minor importance are $n = 5$, 7, 9, etc., where the lower-frequency describing-function loci are of use in the quantitative discussion of the 5th, 7th, 9th, etc., order of sub-harmonic oscillation.

In programming a computer for the analysis of the response of a non-linear element, it is desirable to provide as flexible a programme as possible, so that all the describing functions required may be evaluated with little change of the basic programme. Thus, in eqn. (1) a, b and n are set by the input data tape; ωt and ϕ are varied in a regular manner from 0 to 2π.

The output waveform is evaluated at 96 spot values, and each of these is multiplied by $\cos(\omega t + \phi)$, $-\sin(\omega t + \phi)$, $\cos n\omega t$, and $-\sin n\omega t$ to give the required in-phase and out-of-phase components of the output. When $\omega t \geqslant 2\pi$, so that

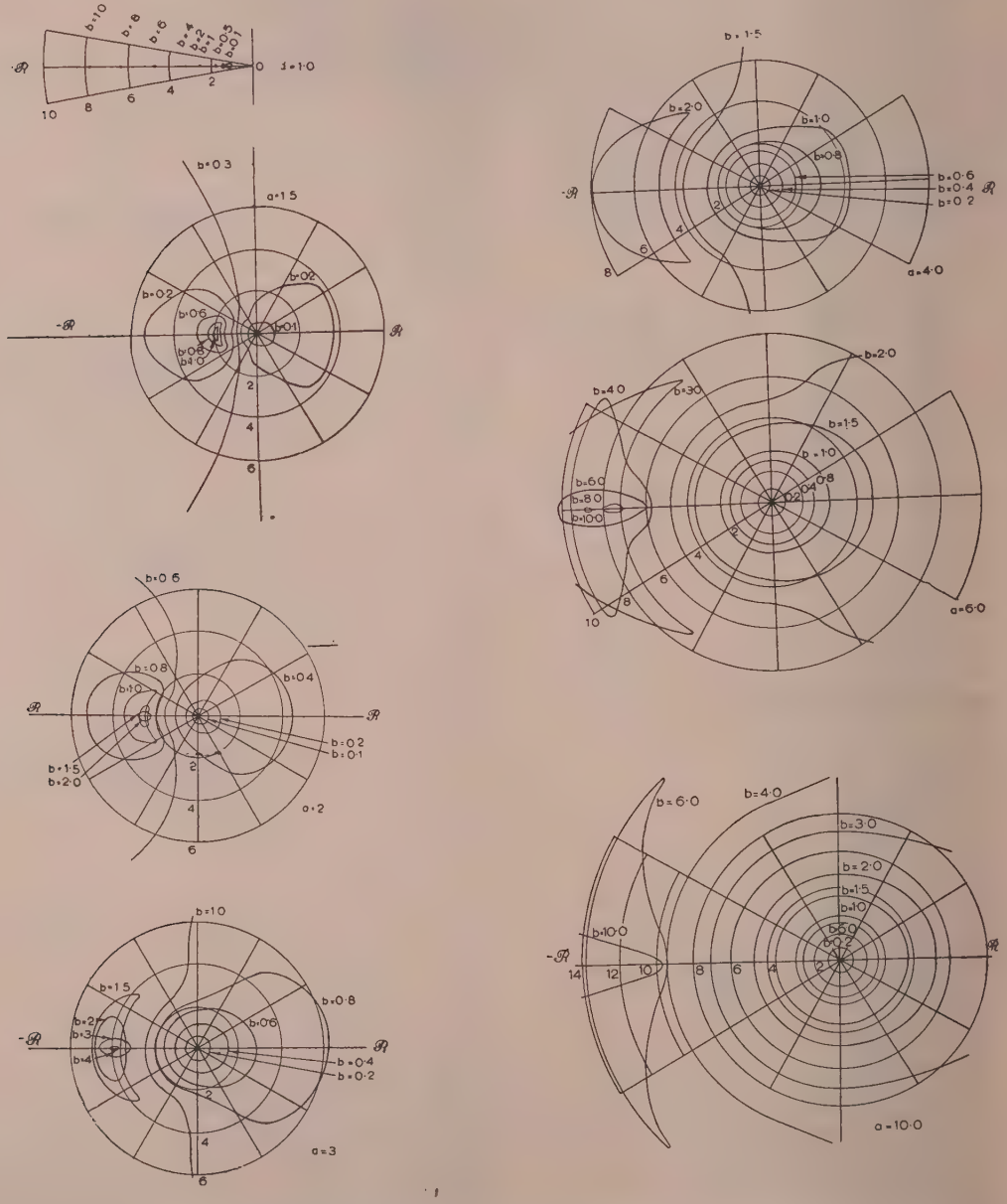


Fig. 4.—High-frequency describing-function loci.

he required number of spot values have been taken, the data s printed out in the columns in this order:

Fraction of harmonic in the output

$\frac{1}{|\text{gain}|}$ of component of frequency $\frac{\omega}{2\pi}$

Tangent of phase-change of this component.

$\frac{1}{|\text{gain}|}$ of component of frequency $\frac{n\omega}{2\pi}$

Tangent of phase-change of this component.

ϕ is then adjusted and the process repeated until $\phi \geqslant 2\pi$, when fresh values of a and b (and possibly n) are read in. A typical set of results is shown in Fig. 1 for $a = 4$, $b = 1$.

This shows that the phase of the low-frequency component is changed by $\pm 13°$ due to the presence of the third harmonic at the input. The phase of the third harmonic, however, varies between the limits $\pm\pi$, as shown by the manner in which the sign of the tangent varies.

Since the describing-function loci are symmetrical about the real axis, the twelve results obtained for values of given a and b determine 24 points on the describing-function locus.

Fig. 2 shows the incremental describing-function loci for a saturation characteristic. For $n = 3$, Fig. 3 gives the low-frequency loci and Fig. 4 those for the component of frequency $3\omega/2\pi$.

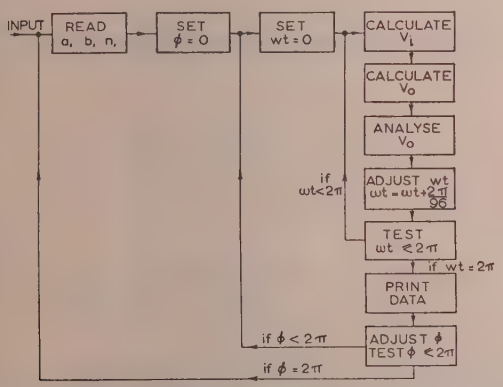


Fig. 5.—Computer operation in analysis of response of non-linear element to input of the form

$v_i = a \sin \omega t + b \sin (n\omega t + \phi)$.

The applications of these general curves are discussed in the following Sections. A "flow sheet" of the operation of the computer is shown in Fig. 5.

(3) STABILITY OF THE FORCED OSCILLATIONS OF A NON-LINEAR FEEDBACK SYSTEM

Consider the non-linear feedback system shown in Fig. 6, where the error detector is followed by a "simple" non-linear element and a linear frequency-dependent network whose response falls with increasing frequency. Let the input be

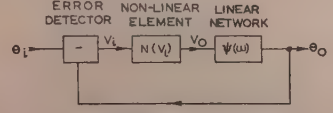


Fig. 6.—Basic system.

written $\theta_i = \hat{\theta}_i \cos(\omega t + \theta)$ and the error signal $v_i = a \cos(\omega t + \phi)$. Where N (v_i) is known, a and $(\phi - \theta)$ may be evaluated by means of a modified describing-function technique.

To investigate the stability of the evaluated steady-state solution, the feedback loop is broken as shown in Fig. 7. The feed-

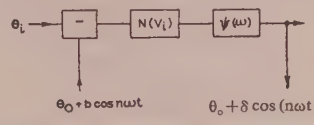


Fig. 7.—Determination of stability of the steady-state response.

back is removed, but a signal θ_0 is injected, equal to the steady-state output of the closed-loop system. In addition to this signal, a small* signal of frequency $n\omega/2\pi$ is injected, giving an output $\delta \cos(n\omega t + \gamma)$. The incremental open-loop gain of the system may be defined as the ratio of the change in output component of frequency $n\omega/2\pi$ to the input amplitude of frequency $n\omega/2\pi$.

As n is varied, both real and imaginary parts of this gain vary, and may be plotted on the complex plane. If this gain locus encloses the $(-1, 0)$ point, any slight disturbance will initiate oscillations of increasing amplitude, so that the steady-state conditions assumed initially are unstable or divergent.

For a linear system, the superposition theorem applies, and hence the incremental open-loop gain is identical to the open-loop gain as measured with $\theta_i = 0$. Hence the steady-state solution is always stable if the system is stable for zero input.

For a non-linear system, the response is a function of the amplitude of the primary input $\hat{\theta}_i$, and of the phase difference between this signal and the additional test signal.

(3.1) The Condition for Instability with Cubic Non-Linearity

As an example of the procedure, consider the cubic non-linearity

$$v_0 = v_i^3$$

with an input to the non-linear element

$$v_i = a \cos(\omega t + \phi) + b \cos n\omega t$$

where b is very small compared with a. The component of v_0 having a frequency $n\omega/2\pi$ is, from Table 1,

$$\frac{3a^2b}{2} \cos n\omega t$$

for all values of n except $n = 1$.†

Thus it can be considered that a component $b \cos n\omega t$ becomes $(3a^2b/2) \cos n\omega t$ after passing through the non-linear element and being superimposed on the steady-state signal.

Thus the incremental gain of the non-linearity is $3a^2/2$ and the incremental open-loop gain is the same as the open-loop gain of the linear portion of the system with the non-linear element replaced by a linear one having a gain of $3a^2/2$. This holds for all frequencies except when $n = 1$ or 3, and is shown in Fig. 8. It follows that if the open-loop frequency-response locus never cuts the negative real axis, this incremental response locus also does not cut this axis. However, when $n = 1$ the component of frequency $n\omega/2\pi$ due to b in the presence of a is

$$\frac{3a^2b}{4}[2 \cos \omega t + \cos(\omega t + 2\phi)] \quad . \quad . \quad . \quad (6)$$

* Theoretically a vanishingly small signal, but in practice the smallest measurable.

† For $n = 3$ there is an additional term in the output of frequency $n\omega/2\pi$ equal to $a^3/4 \cos 3(\omega t + \phi)$, representing distortion of the primary low-frequency signal. Since it is independent of b it is neglected at this stage in the analysis.

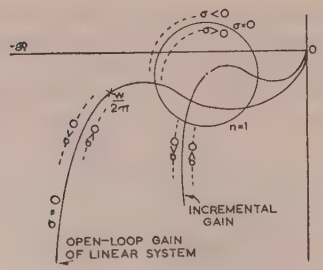


Fig. 8.—Incremental frequency response.
Frequency of primary input = $\omega/2\pi$.

where $b \cos \omega t$ is the incremental input. The ratio of these two gives the incremental gain and is a function of ϕ, an arbitrary phase angle which can assume any value from 0 to 2π. The locus of the incremental gain as ϕ is varied is a circle, with centre at $3a^2/2$ and a radius of $3a^2/4$. The complete incremental Nyquist diagram including the value $n = 1$ is shown in Fig. 8.

Considering an incremental input of the form $b \cos \omega t$, where $b = b_0\varepsilon^{\sigma t}$, it is seen by substitution in eqn. (6) that the corresponding output is $\frac{3a^2b_0\varepsilon^{\sigma t}}{4}[2 \cos \omega t + \cos(\omega t + 2\phi)]$. Hence the gain of the non-linear element is independent of σ. From this result it follows that the variation of the incremental gain of the complete system with σ is entirely due to the frequency-dependent linear system. Thus the stability depends upon whether the incremental open-loop frequency-response locus does or does not enclose the $(-1, 0)$ point.

Consider the response of the system to a sinusoidal signal of angular frequency ω_0. Let $\psi(\omega_0) = x + jy$ be the steady-state response of the linear portion. Hence the incremental gain for

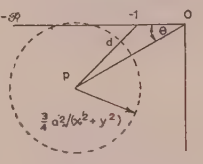


Fig. 9.—Condition for enclosure of the $(-1, 0)$ point.
P is the point $\frac{3}{4}a^2(x + jy)$.

this frequency is a circle with centre at $(3/2)a^2(x + jy)$ and radius, r, $(3/4)a^2(x^2 + y^2)^{\frac{1}{2}}$. The condition for this circle to enclose the $(-1, 0)$ point is derived from Fig. 9.

The distance d of point P from the $(-1, 0)$ point is given by

$$d^2 = 9/4a^4y^2 + 1 + 9/4a^4x^2 - 3a^2x \quad . \quad . \quad (7)$$

The $(-1, 0)$ point is enclosed for $r > d$, i.e.

$$a^4 - (16/9)x/(x^2 + y^2)a^2 + (16/27)/(x^2 + y^2) > 0 \quad . \quad (8)$$

The limit is given by making expression (8) = 0, so that

$$a^2 = \frac{4}{9(x^2 + y^2)}[2x \pm \sqrt{(x^2 - 3y^2)}] \quad . \quad . \quad . \quad (9)$$

Thus, for a real solution, the necessary and sufficient condition is that

$$x^2 \geqslant 3y^2 \quad . \quad . \quad . \quad . \quad . \quad . \quad (10)$$

This gives the condition for an unstable solution as

$$\tan \theta = y/x \leqslant \frac{1}{\sqrt{3}} \quad . \quad . \quad . \quad . \quad , \quad (11)$$

as shown in Fig. 9.

Thus, for the particular non-linearity considered, the sufficie condition for the steady-state solution to be stable is that th open-loop frequency-response locus of the linear system sha never lie within 30° of the negative real axis. This result is vali for any form of linear system, provided that the response fal over the range considered with increasing frequency.

To determine the region of error amplitude for which th system is unstable at a particular frequency, it is convenient t plot the describing function for the incremental response $(-1/G$ From eqn. (6), this is given by

$$-1/G = -\frac{(4)}{(3a^2)}\frac{2 + (\cos 2\phi + j \sin 2\phi)}{5 + 4 \cos 2\phi} \quad . \quad . \quad (12$$

This function is plotted on the complex plane for several value of error amplitude in Fig. 10. The points where these curve

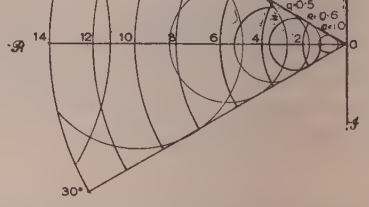


Fig. 10.—Incremental describing function for cubic non-linearity.

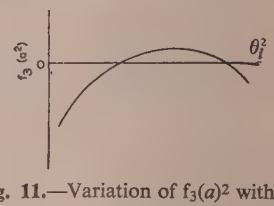


Fig. 11.—Variation of $f_3(a)^2$ with θ_i^2.

cut the response locus of the linear system give the region o instability.

Electronic feedback systems have been employed to verif the theory of the dual-input describing function. Fig. 12(a shows the open-loop frequency-response locus of the linea system in which the cubic non-linearity is inserted. Fig. 12(b shows the experimental and theoretical regions of instability.

(3.2) The Region in which the Jump Effect occurs

The general system with a cubic non-linearity may be readil analysed, mathematically, to give the region of gain and inpu amplitude in which there are three solutions. This region is t be compared with the region previously found in which unstabl solutions occur.

Referring to Fig. 6, let

$$v_i = \theta_i - \theta_0 = a\varepsilon^{j\omega t} \quad . \quad . \quad . \quad . \quad . \quad (13$$

$$\theta_i = \hat{\theta}_i\varepsilon^{j(\omega t - \phi)} \quad . \quad . \quad . \quad . \quad . \quad (14$$

$$\psi(\omega) = x + jy \quad . \quad . \quad . \quad . \quad . \quad (15$$

It follows that

$$\theta_0 = \theta_i - v_i = v_iN(v_i)\psi(\omega)$$

and $$a[1 + N(a)(x + jy)] = \hat{\theta}_i[\cos \phi - j \sin \phi] \quad . \quad (16$$

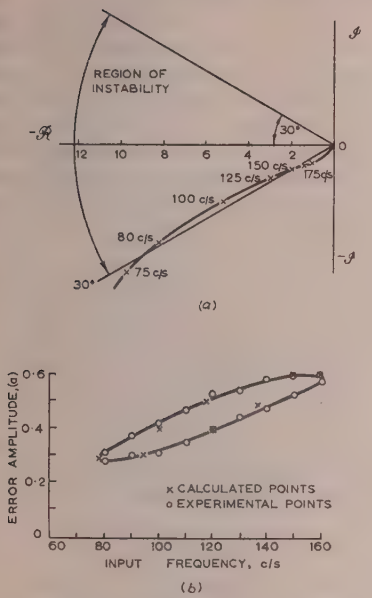


Fig. 12.—Response of the linear system with cubic non-linearity.
(*a*) Open-loop response locus.
(*b*) Instability region: × calculated; ⊙ experimental.

$N(a)$ is real, so that

$$a[1 + xN(a)] = \hat{\theta}_i \cos \phi$$

and

$$ayN(a) = -\hat{\theta}_i \sin \phi$$

Eliminating ϕ,

$$a^2[(x^2 + y^2)N^2(a) + 2xN(a) + 1] = \hat{\theta}_i^2 \quad . \quad . \quad (17)$$

For the cubic non-linearity, $N(a) = \frac{3}{4}a^2$.

Hence

$$a^6 + a^4\left[\frac{8x}{3(x^2 + y^2)}\right] + a^2\left[\frac{16}{9(x^2 + y^2)}\right] - \frac{16\hat{\theta}_i^2}{9(x^2 + y^2)} = 0 \quad . \quad . \quad (18)$$

This is a cubic equation in a^2, and can have three roots only for $x < 0$. When x is negative, there are no negative roots, since the coefficients alternate in sign. Thus the condition for three real positive roots is identical to the condition for three real roots. This condition depends on $\hat{\theta}_i$, x and y.

Sturm's theorem[7] may be applied to this equation to give the number of real roots, as shown in Section 9. It follows, from the analysis, that there are three roots or one, depending on whether

$$f_3(a^2) = -243(x^2 + y^2)2\hat{\theta}_i^4 - 48x(x^2 + 9y^2)\hat{\theta}_i^2 - 64y^2 \gtrless 0 \quad (19)$$

$f_3(a^2)$ is evidently of the form shown in Fig. 11. The limiting condition is that in which there are two equal roots to the equation $f_3(a^2) = 0$, i.e.

$$\left.\begin{aligned} x(x^2 + 9y^2)2 - 27y^2(x^2 + y^2)2 &= 0 \\ (x^2 - 3y^2)^3 &= 0 \end{aligned}\right\} \quad . \quad . \quad (20)$$

or

$$\frac{y}{x} = \frac{1}{\sqrt{3}} \quad . \quad . \quad . \quad . \quad . \quad . \quad (21)$$

For $x^2 > 3y^2$ there are two solutions to the equation $f_3(a^2) = 0$. Thus this analysis shows that there can be three steady-state values of error amplitude for a given input amplitude only if the phase change of the linear element is between 150° and 210°. This result is in agreement with the preceding theory of the region of an unstable solution.

(3.3) The Stability of a System incorporating Saturation

To derive the region of instability for a saturating system subject to sinusoidal input, the incremental describing-function loci of Fig. 2 are superimposed on the open-loop frequency-response locus of the linear system. An unstable solution is predicted if the describing-function locus for the particular amplitude of signal considered, a, encloses the point on the complex plane representing the open-loop gain of the linear system at the forcing frequency.

From these loci it follows that the region of unstable solutions can be defined in terms of input frequency and amplitude of signal applied to the non-linear element.

The describing-function loci all lie in a certain region on the complex plane, as seen from Fig. 2, and if the open-loop response for the input frequency considered lies outside this region, no instability will be observed at this particular frequency, for any input amplitude.

Consider now the case where the point representing the open-loop response at the input frequency lies inside this region. For $a \leqslant 1$, the non-linear element behaves linearly, so that the incremental gain is unity and the describing function is a vector from the origin to the $(-1, 0)$ point. Hence, for an unstable solution, the frequency-response locus must enclose the $(-1, 0)$ point. As the amplitude of the input to the non-linearity increases, the incremental describing function expands and will touch the point representing the open-loop response of the linear system at the input frequency considered. This represents an overall open-loop gain of unity for an incremental signal, so that any incremental signal is maintained (theoretically) with constant amplitude.

When the amplitude of the signal applied to the non-linear element is such that the relevant describing-function locus encloses the point representing the response at input frequency, any incremental signal is regenerated with increasing amplitude. The system therefore diverges from the state initially postulated. This implies that the amplitude of the signal, v_i, initially considered is, in fact, unstable.

It is evident that as the amplitude of the signal applied to the saturating non-linear element is increased further, the incremental gain decreases and the incremental describing function increases in magnitude. At some value of a the describing function no longer encloses the point representing the gain of the linear system, so that this value of a represents a stable steady-state solution.

The range of unstable amplitudes of error signal is found in this way for several input frequencies. The practical systems considered simulate a second-order r.p.c. servo mechanism, having open-loop frequency-response loci shown in Fig. 13(*a*). Superimposing the amplitude-dependent loci of Fig. 2 on these frequency-dependent curves gives the regions of a and frequency corresponding to unstable solutions. Experimental results agree well with the theoretical analysis [Fig. 13(*b*)].

The stability of the third solution has been investigated experimentally by means of the apparatus shown in Fig. 14. It is seen that the feedback loop is opened, and a further signal θ_0' injected equal to the predicted steady-state value of θ_0. In this state the trace on the cathode-ray oscillograph is a straight line making an angle of 45° with the horizontal. The switch is then operated, removing θ_0' and closing the feedback loop. For a stable solution, θ_0 remains unchanged. In every case in which three solutions are predicted, the intermediate amplitude of θ_0

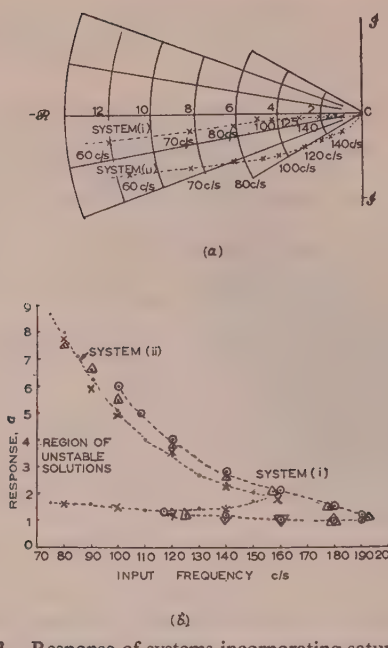


Fig. 13.—Response of systems incorporating saturation.
(*a*) Open-loop response loci.
(*b*) Instability regions: × △ calculated; ● ○ experimental.

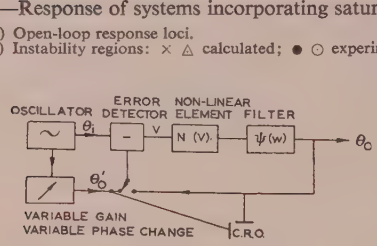


Fig. 14.—Experimental investigation of stability of a given solution.

is found to be unstable. When the loop is closed and the initial conditions correspond to an unstable solution, either stable solution can result as the steady state.

(4) THE AMPLITUDE STABILITY OF CONDITIONALLY STABLE SYSTEMS

A conditionally stable system in the Nyquist sense for purely linear elements has a frequency range in which the open-loop gain is greater than unity whilst the angle of lag is greater than 180°. An example is shown in Fig. 15. It is well known that

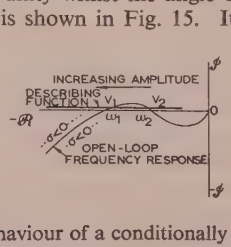


Fig. 15.—Behaviour of a conditionally stable system.

owing to the presence of non-linearities such as saturation such systems are susceptible to continuous oscillation if excited by suitable forms of input signal. This oscillation is maintained after the input signal is removed.

The describing-function technique enables the amplitude and frequency of the free oscillations to be determined.[1] The describing function for the non-linearity is plotted on the complex plane, and the point at which this intersects the open-loo frequency-response locus of the linear system gives the operatin position. In Fig. 15 a hypothetical describing function on th negative real axis is drawn cutting the Nyquist diagram at point corresponding to frequencies ω_2 and ω_1. These occur at ampli tudes v_2 and v_1 respectively. For the direction of increasin amplitude indicated, the point (v_2, ω_2) is unstable and divergent the system either decreasing amplitude and ceasing oscillatio (moving to the right) or increasing amplitude and frequenc transiently until the point (v_1, ω_1) is reached, which is a stabl oscillation.

It is known in practice, however, that most forms of inpu will excite this particular oscillation if the amplitude is suffi ciently large. The dual-input describing-function techniqu enables the critical amplitude of a sinusoidal input signal o any frequency to be determined. This is achieved in the sam manner as that described in Section 3. Suppose the input signa of variable amplitude but fixed frequency $\omega/2\pi$ is applied Then the incremental Nyquist diagram can be obtained for small signal superimposed on the steady-state conditions, i.e an additional signal of frequency $n\omega$, where n is varied from to ∞.

For the cubic non-linearity already considered with a signa

$$v_i = a\cos(\omega t + \phi) + b\cos n\omega t$$

appearing at the input to the non-linear element, the gain of th element for the component of frequency $n\omega/2\pi$ is $3a^2/2$. Thu the incremental diagram is similar to the open-loop frequenc response of the linear portion multiplied by the factor $3a^2/2$. Th value of a which causes oscillation can be ascertained and i independent of frequency. An important result is that th

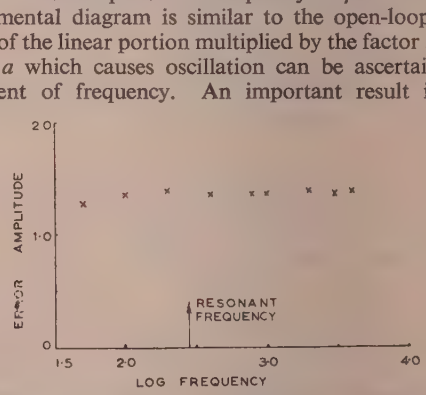


Fig. 16.—Experimental verification of critical input.

critical amplitude of input signal of any frequency is the sam when measured at the input to the non-linear element. This wa verified experimentally as shown in Fig. 16.

(5) SYNCHRONIZED OSCILLATIONS

It has been shown that the response of a non-linear elemen to a sinusoidal signal can be modified in amplitude and phase b the presence of a further sinusoid of simply related frequency This effect modifies the loop gain of the system, and may mak the closed loop continuously oscillatory, the frequency o oscillation being simply related to the input frequency.

The dual-input describing-function technique enables th occurrence of these oscillations to be predicted quantitatively The procedure is similar for both high- and low-frequenc forced oscillations. The two cases are discussed separately t emphasize the differences.

(5.1) Sub-Harmonic Oscillations

To investigate a sub-harmonic whose frequency is $1/n$th tha of the primary input, the signal $b\cos n\omega t$ [eqn. (1)] is regarde as predetermined, and the describing function for the signa

$a \cos(\omega t + \phi)$ is superimposed on the open-loop frequency-response locus of the linear system. The value of a, such that the describing function cuts the frequency-response locus for a frequency $\omega/2\pi$, is the amplitude corresponding to an overall open-loop gain of unity. Hence this gives a possible steady-state amplitude of the resultant sub-harmonic oscillation. For both non-linearities considered, the phase-change for the low-frequency component is small for small amplitudes of this component. As a result, the sub-harmonic oscillations are not self-starting but must be initiated by an additional impulse.

Sub-harmonics in a system with cubic non-linearity have been discussed previously.[6] It has been shown that sub-harmonics can occur whenever the open-loop frequency-response locus lies within $\pm 21°$ of the negative real axis. In this system only the third sub-harmonic can occur.

Considering the saturation characteristic defined in Section 2.2, the amplitude and frequency of any possible sub-harmonic oscillation of order 1/3 is found as follows.

Initially, the amplitude of the signal of input frequency applied to the non-linear element is found by standard techniques.[3] Since the frequency of the sub-harmonic oscillation is of the order of the natural frequency of the linear system, the system response is small at the input frequency considered. Hence, as a first approximation, the amplitude of the error signal may be derived by putting $\theta_0 = 0$. For the system to oscillate continuously at a frequency one-third that of the input the overall open-loop gain for this low-frequency component must be unity. Thus the conditions for continuous oscillation may be evaluated by means of a describing-function technique.

The input to the non-linear element is of the form

$$v_i = a \cos(\omega t + \phi) + b \cos 3\omega t$$

where the signal $b \cos 3\omega t$ is the primary input, of known amplitude. The loci of Fig. 3 show the response of the saturation-type non-linearity to a sinusoidal signal of amplitude a in the presence of a signal of three times this frequency and of amplitude b. The revelant loci of Fig. 3 are determined from a knowledge of the operating value of b. The set of loci for this particular value of b show the effect of varying a, the amplitude of the sub-harmonic oscillation.

These amplitude-dependent loci are superimposed on the open-loop frequency-response locus of the linear system. The point at which they pass through the point representing the frequency response at the sub-harmonic frequency determines a, the amplitude of the sub-harmonic oscillation.

For each value of input frequency the range of b in which sub-harmonic resonance can occur is evaluated, and the corresponding sub-harmonic amplitude is found from the loci of Fig. 3.

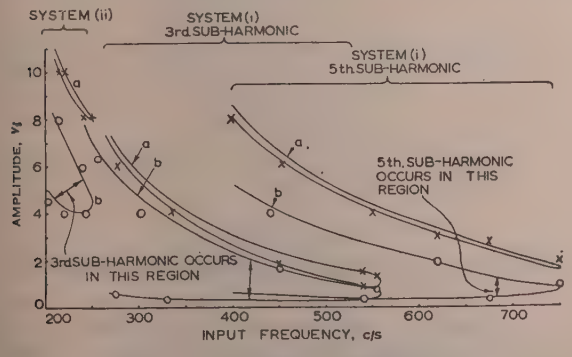


Fig. 17.—Occurrence of sub-harmonics.

× Calculated values of a.
⊙ Calculated values of b.

Experimental work has here again made use of the second-order systems whose open-loop frequency-response loci are shown in Fig. 13(*a*). The range of b for which sub-harmonic resonance occurs and the resultant amplitude of the oscillation have been determined over a range of input frequencies, and the experimental and theoretical results are shown in Fig. 17. This gives the regions of input for the observation of the third sub-harmonic in both systems, and for the fifth sub-harmonic in the lightly damped system.

(5.2) Stability of the Sub-Harmonic Oscillations

In general, it is found that there are two values of a which satisfy the conditions for continuous oscillation shown in Fig. 18.

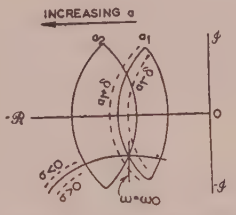


Fig. 18.—Stability of sub-harmonic oscillations.

The intermediate value a_1, where $0 < a_1 < a_2$, is always unstable. This is shown by considering the effect of a small variation in a. Since the describing-function locus refers to one particular frequency, say ω_0, equal to one-third the input frequency in this case, the operating point must always lie along the line $\omega = \omega_0$. Hence a small increase in amplitude from $a = a_1$ gives a solution $\omega = \omega_0$, $\sigma > 0$, leading to a further increase in amplitude. Similarly a small decrease in amplitude causes a cumulative decay in a. It also follows that the oscillation of amplitude a_2 is stable as regards small changes of amplitude.

Thus the occurrence of three isochronous solutions and of the sub-harmonic oscillations illustrates that stable and unstable solutions occur alternately and disappear in pairs.

(5.3) Harmonic Oscillations

When estimating the amount of third harmonic present in the response of a system to a sinusoidal input, it is usual to assume that the error signal is sinusoidal. The output of the non-linear element is then analysed into its Fourier components. This method is accurate only when the open-loop gain of the system is much less than unity at the frequency of the harmonic.

Since the harmonic components constitute forced oscillation

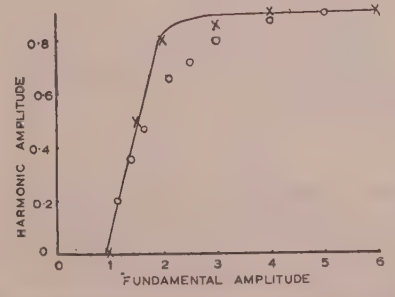


Fig. 19.—Harmonic content of error signal for system (ii) of Fig. 13(*a*).

⊙ Calculated values, ignoring effect of third harmonic.
× Calculated values from dual-input describing function.
The curve is plotted from experimental observations.
Input frequency, 50 c/s.

of the system, the loop gain for each harmonic component must be unity. Thus again, the describing-function technique can be used to find the conditions for continuous oscillation. As in the case of sub-harmonic oscillations (Section 5.1) the input to the non-linear element is assumed to be of the form

$$v_i = a \cos(\omega t + \phi) + b \cos 3\omega t$$

In this case, the component $a \cos(\omega t + \phi)$ is due to the primary input and is regarded as predetermined. For unity overall loop-gain of the harmonic component, the amplitude-dependent describing function for this component cuts the frequency-response characteristic of the linear system at a frequency three times the input frequency. Thus the variation of b with a and input frequency is determined.

For saturation, the relevant loci are shown in Fig. 4. Experimental and theoretical results for system (ii) in Fig. 13(*a*) are given in Fig. 19. For comparison, this Figure also shows the theoretical harmonic content when the effect of this harmonic on the output of the non-linear element is neglected.

(6) EFFECT OF THE THIRD HARMONIC

The previous Section described a method of determining the harmonic content in the response to a given input. This harmonic content being known, it is possible to evaluate more accurately the response at fundamental frequency.

The closed-loop frequency response is found by a method introduced by Hammond* using the open-loop frequency response in connection with the describing function ($b = 0$). For known values of b and ϕ the dual-input describing function [Fig. (3)] is employed.

For the systems considered it is found that the influence of the third harmonic is negligible. The gain is practically unchanged, but the lag of the output is increased by a few degrees. This is the most satisfactory justification for the assumption of a sinusoidal error signal; 50% of third-harmonic amplitude at low frequencies has very little effect on the closed-loop gain.

Another application of a similar nature occurs in the consideration of a sinusoidal signal passing through two non-linear elements separated by a frequency-responsive network. The output of the first non-linear element may in many cases be approximately expressed in the form

$$a \cos(\omega t + \phi_1) + b \cos(3\omega t + \phi_2)$$

so that the input to the second non-linear element is of the form

$$a_1 \cos(\omega t + \phi) + b \cos 3\omega t$$

The dual-input describing function enables the gain and phase change of the component of frequency $\omega/2\pi$ to be determined.

Work is proceeding on the determination of the stability and frequency response of the Serme system,[11] which consists essentially of a non-linear element of the form $v_0 = (v_i)^{\frac{1}{2}}$ and a marked saturation characteristic, the non-linear elements being separated by a stabilizing network.

(7) CONCLUSION

A technique has been presented which shows how the response of a non-linear element to a complex input can be employed to predict the behaviour of a non-linear feedback system. Many phenomena observed in non-linear systems can be explained quantitatively by its use.

It is shown that the open-loop frequency response of the linear system is invaluable for deducing the response of the non-linear system. From this response it is possible to examine the stability of the response of the non-linear system to a sinusoidal input.

For certain systems it is found that there is a region of input amplitude and frequency in which the response, as given by steady-state analysis is, in fact, unstable. This region of instability is found to coincide with the region in which the jump effect occurs.

The technique presented also gives a physical explanation of the mechanism of sub-harmonic generation and of the mechanism by which a conditionally stable system can be set into permanent oscillation. Graphical methods evaluate the region of input amplitude and frequency in which these phenomena occur.

The harmonic content in the response to a sinusoidal input is readily found by the same technique.

For particular systems, experimental results verify the theoretical treatment.

* Reference 4, p. 489.

(8) REFERENCES

(1) KOCHENBURGER, R. J.: "A Frequency Response Method for Analysing and Synthesising Contactor Servomechanisms," *Transactions of the American I.E.E.*, 1950, **69**, Part I, p. 270.

(2) JOHNSON, E. C.: "Sinusoidal Analysis of Feedback Control Systems containing Non-Linear Elements," *ibid.*, 1952, **71**, p. 169.

(3) WEST, J. C., and DOUCE, J. L.: "The Frequency Response of a certain class of Non-Linear Feedback Systems," *British Journal of Applied Physics*, 1954, **5**, p. 204.

(4) WEST, J. C., and NIKIFORUK, P.: "The Behaviour of a Remote-Position-Control Servo mechanism with Hard Spring Non-Linear Characteristics," *Proceedings I.E.E.*, Paper No. 1621 M, March, 1954 (**101**, Part II, p. 481).

(5) STOKER, J. J.: "Non-Linear Vibrations" (New York: Interscience Publishers, 1950).

(6) WEST, J. C., and DOUCE, J. L.: "The Mechanism of Sub-Harmonic Generation in a Feedback System," *Proceedings I.E.E.*, Paper No. 1693 M, July, 1954 (**102** B, p 569).

(7) TURNBULL, H. W.: "Theory of Equations" (Oliver and Boyd, 1942).

(8) CHERRY, E. C., and MILLAR, W.: "Automatic and Manual Control" (London: Butterworths' Scientific Publications Ltd., 1952), p. 263.

(9) WEST, J. C.: "The Nyquist Criterion of Stability," *Electronic Engineering*, 1950, **22**, p. 169.

(10) WEST, J. C.: "Text Book of Servomechanisms" (English Universities Press, 1954).

(11) WEST, J. C., DOUCE, J. L., and NAYLOR, R.: "The Effects of the Addition of Some Non-Linear Elements on the Transient Performance of a Simple R.P.C. System possessing Torque Limitation," *Proceedings I.E.E.*, Paper No. 1549 M, August, 1953 (**101**, Part II, p. 156).

(12) TUSTIN, A.: "A Method of Analysing the Behaviour of Linear Systems in Terms of Time Series," *Journal I.E.E.*, 1947, **94**, Part IIA, p. 130.

(13) MINORSKY, N.: "Introduction to Non-Linear Mechanics" (Edwards, Ann Arbor, 1947).

(9) APPENDIX

The condition for three real roots of the cubic equation

$$f(\lambda) = \lambda^3 + b\lambda^2 + c\lambda + d = 0 \quad . \quad . \quad . \quad (22$$

may conveniently be evaluated from Sturm's theorem. The first step is to form the derivative $f'(\lambda)$ of $f(\lambda)$. This derivative involves terms of degree λ^2 and lower. Two further functions $f_2(\lambda)$ and $f_3(\lambda)$ are required. $f_2(\lambda)$ is defined as the remainde

fter dividing $f(\lambda)$ by $f'(\lambda)$ and $f_3(\lambda)$ is the remainder after dividing $f'(\lambda)$ by $f_2(\lambda)$. $f_3(\lambda)$ is independent of λ.

Table 2 is then prepared thus:

Table 2

λ	$f(\lambda)$	$f'(\lambda)$	$f_2(\lambda)$	$f_3(\lambda)$	No. of changes of sign
λ_1	+	+	+	+	N_1
	−	−	−	−	
λ_2	+	+	+	+	N_2
	−	−	−	−	
λ_3	+	+	+	+	N_3
	−	−	−	−	

the sign of each function being determined for the given value of λ. The end column gives the number of changes of sign along a horizontal row, from $f(\lambda)$ to $f_3(\lambda)$. The maximum number is 3 in this case.

Sturm's theorem states that the number of real roots between λ_1 and λ_2 is equal to the difference in the number of changes of sign in the tabulated functions for $\lambda = \lambda_1$ and $\lambda = \lambda_2$, i.e. $N_1 - N_2$.

In the case considered, where $\lambda = a^2$, we have, from eqn. 18,

$$f(\lambda) = \lambda^3 + \frac{8}{3}\frac{x}{x^2+y^2}\lambda^2 + \frac{16}{9}\frac{1}{x^2+y^2}\lambda - \frac{16\hat{\theta}_i^2}{9(x^2+y^2)}$$

$$f'(\lambda) = 3\lambda^2 + \frac{16}{3}\frac{x}{x^2+y^2}\lambda + \frac{16}{9(x^2+y^2)}$$

$$f_2(\lambda) = (2x^2 - 6\lambda^2)\lambda + \left[9(x^2+y^2)\hat{\theta}_i^2 + \frac{8x}{3}\right]$$

$$f_3(\lambda) = -243(x^2+y^2)^2\hat{\theta}_i^4 - 48x(x^2+9y^2)\hat{\theta}_i^2 - 64y^2$$

There are two cases to be considered for $\lambda \to \pm\infty$:

(i) $f_2(\lambda) = -\operatorname{sgn}\lambda$ i.e. $x^2 < 3y^2$

(ii) $f_2(\lambda) = \operatorname{sgn}\lambda$ $x^2 > 3y^2$

Case (i): $f_3(\lambda)$ is always negative in this case. The Table will appear thus:

λ	$f(\lambda)$	$f'(\lambda)$	$f_2(\lambda)$	$f_3(\lambda)$	Changes of sign
$-\infty$	−	+	−	−	2
$+\infty$	+	+	+	−	1

Thus there is always one and only one real root for $x^2 < 3y^2$.

Case (ii). $f_3(\lambda)$ may now be positive or negative. For $f_3(\lambda) < 0$ the Table will be

λ	$f(\lambda)$	$f'(\lambda)$	$f_2(\lambda)$	$f_3(\lambda)$	Changes of sign
$-\infty$	−	+	−	−	2
$+\infty$	+	+	+	−	1

i.e. there is always one and only one real root.

For $f_3(\lambda) > 0$ the Table will be

λ	$f(\lambda)$	$f'(\lambda)$	$f_2(\lambda)$	$f_3(\lambda)$	Changes of sign
$-\infty$	−	+	−	+	3
$+\infty$	+	+	+	+	0

Thus the condition for three real positive roots is that $x^2 > 3y^2$ and $f_3(\lambda) > 0$.

Eliminating $\hat{\theta}_i^2$ from $f_3(\lambda)$, as in the text, shows that three real roots are possible provided that $x^2 > 3y^2$.

Thus for three solutions at a given input frequency, for some input amplitude, the necessary and sufficient conditions are that:

(i) $x < 0$

(ii) $x^2 > 3y^2$

These conditions are identical with those given by graphical construction from the Nyquist diagram.

DISCUSSION BEFORE THE MEASUREMENT AND CONTROL SECTION, 13TH MARCH, 1956

Mr. J. Bell: Many years ago at the Admiralty, I was engaged on the stabilization of equipment on ships. One of the problems with which we were concerned was trying to make equipment follow the motion of the ship and, in effect, point to the horizon. For example, a searchlight was required to point at the horizon, instead of waving about in the sky. This was, of course, before radar equipment was available, and the results obtained when the servo mechanism saturated remind me very much of what the authors have found.

Suppose we imagine that the motion of the ship has a gradually increasing amplitude and constant frequency, as shown in Fig. A (it never is like this, actually). The equipment follows the roll, but it is not fully compensated, and so it lags behind a little, as shown by the dotted line changing over at the peak of the sine wave. As the roll velocity increases there comes a point of saturation, or maximum speed of the following equipment, and a discontinuity occurs at 1. The lag increases out of all proportion to the difference in speed between input and output, the following curve takes on a pointed or triangular character and very appreciable phase change results.

It appears that the 'jump' or change in amplitude for an increasing frequency of input, as mentioned by the authors, will probably occur between the points 1 and 2, after which the servo mechanism is continuous by operating at full speed.

It may be noted that the phase displacement between the sine wave and the 'following' curve will approach 90° (lagging) as the amplitude of the former approaches infinity. With decreasing amplitude the 'following' curve also lags if the motion as shown

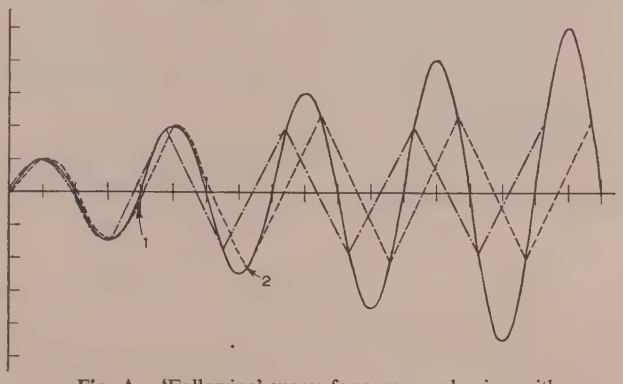


Fig. A.—'Following' curves for servo mechanism with limited speed.

——— Input, which is sinusoidal with variable amplitude.
– – – – 'Following' curve for increasing amplitude.
—·—·— 'Following' curve for decreasing amplitude.

is considered as going from right to left of the diagram for decreasing amplitude.

While this Figure is drawn for constant frequency and variable amplitude, it may be somewhat comparable with the case cited by the authors of constant amplitude and variable frequency.

Drs. J. C. West, J. L. Douce and **R. K. Livesley** (*in reply*): The practical example quoted by Mr. Bell shows the 'jump' effect in a very striking manner. Interest has recently been concentrated on torque-limited systems, so that the output, for a sinusoidal input, always approximates to a sine wave. The velocity signal, in these systems, very closely agrees with the waveforms of Fig. A.

Analysis of the velocity-limited system is identical to that of the torque-limited case. In particular, the 'jump' effect can occur whenever the Nyquist diagram of the linear system cuts the describing-function loci of Fig. 2.

DISCUSSION ON 'ARTIFICIAL REVERBERATION'*

NORTH-EASTERN RADIO AND MEASUREMENTS GROUP, AT NEWCASTLE UPON TYNE, 5TH DECEMBER, 1955

Mr. J. Queen: Research on many types of delay channel has been mentioned, but no reference is made to the possible use of reactive artificial telephone lines or filter networks.

The discussion on methods of testing reverberation implies that steady-state tones are used. The nature of testing equipment has been practically omitted from the paper, and I would like to know how the reverberation characteristics of systems are measured in research investigations.

Some difficulty is experienced in understanding how the effect of resonance could be avoided in the design of tube and room reverberation systems, particularly the resonances of microphones and loudspeakers.

In the case of multi-channel reverberation systems, the effect of inter-channel crosstalk seems to be a factor which must be considered, but it has not been mentioned in the paper. Each channel is provided with an attenuator, and I would like to know whether these could not be dispensed with, and the necessary adjustment of attenuation effected by manipulation of channel reflecting surfaces and selection of material used for this purpose.

When echo rooms are used, it seems possible that they tend to have better response for tones of low frequency than for those of the higher frequency.

Dr. P. E. Axon, and **Messrs. C. L. S. Gilford** and **D. E. L. Shorter** (*in reply*): It is, of course, possible to obtain delays from reactive networks. In this connection a study was made of lattice networks giving group delays which reach a maximum value over a limited frequency range. However, the maximum delay which can be achieved from a single-section network using easily available components is of the order of 2 millisec. Therefore, to obtain the required delay of up to 2 or 4 sec over the whole audio-frequency band would require a prohibitively large number of sections.

Steady-state testing of reverberation systems is an essential and convenient method in development. However, final tests to examine the behaviour of a system with short pulses of tone and various types of programme, and to determine its decay time by direct measurement, are also desirable. For the latter use is made of studio reverberation-testing apparatus which has been described elsewhere.* A simplified and more compact version of this apparatus has since been constructed for routine checking of operational equipment.

When high-quality loudspeakers and microphones are used the defects in performance due to mechanical resonances are unimportant in comparison with the irregularities in the frequency characteristics of tubes and reverberation rooms.

In the case of multi-channel reverberation systems generally inter-channel crosstalk has to be considered, but in our particular case the only multi-channel system tried was the experimental three-tube combination referred to in Section 7.2, in which crosstalk was avoided by making each of the channels a mechanically separate unit.

Adjustment of the attenuation in an acoustic delay tube by the manipulation of internal reflecting surfaces is not only inconvenient but inadmissible when feedback is applied, since the resulting interference effects will lead to a condition similar to that obtaining in the multi-path system envisaged in Section 6.3. In the magnetic system there is, of course, no alternative to electrical adjustment of attenuation.

It is true that echo rooms are normally 'bass heavy'. It has been universal in the past to make echo rooms as reverberant as possible, and the use of any form of sound absorber has been avoided. The consequence is that reverberation time tends to be longer in the bass than in the upper-frequency range, where there is appreciable incidental absorption. More recently it has become the practice in the B.B.C. to introduce a small amount of low-frequency absorption to give a better characteristic.

* Axon, P. E., Gilford, C. L. S., and Shorter, D. E. L.: Paper No. 1796 R, February, 1955 (see 102 B, p. 624).

* Somerville, T., and Gilford, C. L. S.: 'Cathode-Ray Displays of Acoustic Phenomena and their Interpretation', *B.B.C. Quarterly*, 1952, 7, p. 1.

621.372.4/.5: 621.391: 621.3.018.78

Paper No. 2098 R
July 1956

THE EFFECT UPON PULSE RESPONSE OF DELAY VARIATION AT LOW AND MIDDLE FREQUENCIES

with Special Application to Vestigial-Sideband Systems

By M. V. CALLENDAR, M.A., Associate Member.

(*The paper was first received 8th November, 1955, and in revised form 8th February, 1956.*)

SUMMARY

Calculations are given for the magnitude and form of the distortion introduced into a square wave by a network or system which exhibits uniform transmission except for increasing (or decreasing) phase delay in the low-mid-frequency region. The fractional peak distortion is found to be equal to twice the area under the curve relating T_n to frequency, where T_n is the delay relative to that at high frequencies. The waveform of the distortion is given for several simple shapes of curve for T_n. This distortion is especially characteristic of vestigial-sideband systems, and occurs in television as a 'pre-shoot' before a transition and as a smear (in principle equal, but opposite, to the pre-shoot) after it.

(1) INTRODUCTION

The relations between non-uniformity in the steady state (amplitude/frequency and phase-delay/frequency) characteristics and the consequent distortion of pulse or square-wave signals, are, in general, far from simple. To calculate the pulse response for any particular case by Fourier or Laplace transform methods is a matter of considerable labour, and few useful generalizations can be obtained in this way. As a result, some writers have used the simpler relations which apply when the delay is assumed to be uniform (or the phase/frequency relation is linear); however, this procedure leads to large errors in most practical cases, except, of course, where special phase-correcting circuits have been introduced.

The alternative procedure of considering only the delay errors is no better, so far as the usual cut-off distortion is concerned, but there is one case where the effects of non-uniform amplitude characteristics are often negligible compared with those of non-uniform delay. This is the important case of a vestigial-sideband system, where a large distortion of pulse waveforms often occurs as a result of the increased delay through the system at low and middle video frequencies which arises from the tuned circuits in the receiver and from the vestigial-sideband filter in the transmitter.

Although it has been generally realized that the 'pre-shoot' which is often a noticeable feature of television pictures is caused by delay errors, little attention appears to have been given to the effect until recently,[1,2] and no direct relations connecting the amplitude and waveform of the distortion with the delay characteristic appear to have been published. We shall therefore attempt an analysis of the distortion which results when a square wave is applied to a network (or system) having a delay characteristic of the type in question, as illustrated in Fig. 1. The distortion for a low-frequency square wave is, of course, the same as that for a step or for a rectangular pulse of adequate length and low repetition frequency; since the distortion is not connected with the high-frequency components of the signal, there is no need to complicate matters by making calculations for frequency-limited (e.g. sine-squared-pulse) waveforms.

(2) THEORY

For a square wave having a double amplitude of unity

$$V = \frac{2}{\pi}\left(\cos \omega_1 t - \frac{1}{3}\cos 3\omega_1 t + \frac{1}{5}\cos 5\omega_1 t - \ldots\right) \quad . \quad (1)$$

This is applied to a network which leaves the amplitudes of all components unaltered and introduces a phase delay which equals T_c at high frequencies but increases at mid and lower frequencies, so that the delays for the first, third, nth, etc., harmonic components of the square wave are $(T_c + T_1)$, $(T_c + T_3)$, $(T_c + T_n)$, etc. (see Fig. 1). Any practical network

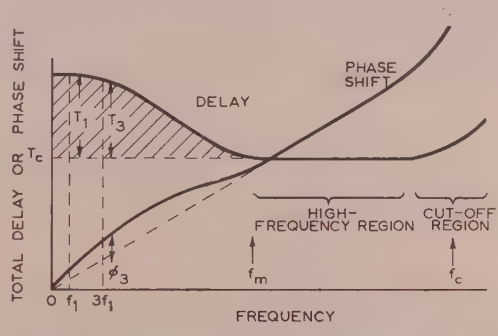


Fig. 1.—Characteristics for delay and phase shift.

T_1, T_3, T_n represents the delay relative to that, T_c, at a high frequency, and the phase differences ϕ_1, ϕ_3, etc., correspond to T_1, T_3, etc. The area shaded is A_T.

will also show a change of delay at very high frequencies near cut-off, but the distortion resulting from this can be treated separately from the low- and mid-frequency distortion considered here, provided only that there is a reasonably wide region over which the delay is substantially constant ($=T_c$). T_1, T_3, T_n, etc., will be termed the 'relative delay' to avoid confusion with the 'delay error', ΔT, which is normally measured relative to the delay at a very low frequency.

The distortion wave, or the difference between the output wave and the input wave delayed by T_c, is given by

$$\Delta V = \frac{2}{\pi}\Big\{\cos \omega_1 t - \cos(\omega_1 t - \omega_1 T_1) - \frac{1}{3}[\cos 3\omega_1 t - \cos(3\omega_1 t - 3\omega_1 T_3)] + \ldots \pm \frac{1}{n}[\cos n\omega_1 t - \cos(n\omega_1 t - n\omega_1 T_n)]\Big\} \quad . \quad (2)$$

where n is an odd integer.

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.

Mr. Callendar is with E. K. Cole, Ltd.

Transforming and substituting $2\pi/t_1$ for ω_1 gives

$$\Delta V = \frac{4}{\pi}\left[\sin\pi\left(\frac{2t-T_1}{t_1}\right)\sin\pi\frac{T_1}{t_1} - \frac{1}{3}\sin 3\pi\left(\frac{2t-T_3}{t_1}\right)\sin 3\pi\frac{T_3}{t_1} + \ldots \pm\frac{1}{n}\sin n\pi\left(\frac{2t-T_n}{t_1}\right)\sin n\pi\frac{T_n}{t_1}\right] \quad . \quad (3)$$

Now if $T_n \ll t_1/n$, where n is the order of the harmonic component, we have

$$\sin n\pi T_n/t_1 = n\pi T_n/t_1$$

and $$\sin n\pi(2t-T_n)/t_1 = \pm 1 \text{ at } t = t_1/4$$

The error in both cases is less than 5% if $T_n < 0{\cdot}1t_1/n$ or $\phi_n = T_n\omega_n < 36°$ for all components. Thus, with this proviso, all the harmonic components will be effectively in phase at $t = t_1/4$, showing that the peaks of the distortion wave occur at the transitions of the output, and that this peak distortion is given by

$$\Delta V_m = 4(T_1 + T_3 + T_5 + \ldots)/t_1 = 4f_1(T_1 + T_3 + T_5 + \ldots) \quad . \quad . \quad . \quad . \quad (4)$$

This formula does not necessarily imply an increase in ΔV_m with increase in the square-wave frequency f_1, since the number of components delayed is thereby decreased. If ϕ is greater than 36° but still less than, say, 90°, the true peak distortion will be a little less than that given above, and the peak will occur just after $t = t_1/4$.

This distortion wave ΔV [Fig. 2(*a*)] is visible as a pre-shoot

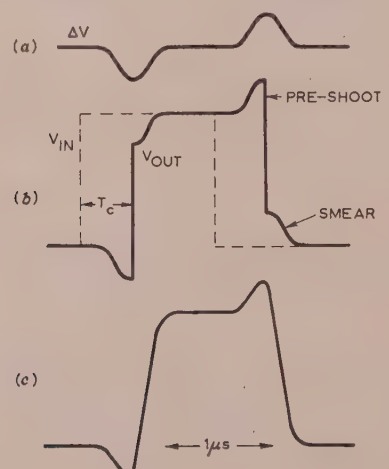


Fig. 2.—Typical distortion of a pulse by low-mid-frequency delay.
(*a*) Distortion wave.
(*b*) Distorted output (as calculated).
(*c*) Output when heavy cut-off distortion is also present.

immediately before the transition and a smear of equal amplitude and identical (but inverted) form after it [Fig. 2(*b*)]. The distortion actually seen on practical pulse tests represents ΔV plus the cut-off distortion, and the latter may modify the form and reduce the apparent amplitude of the smear, as shown in Fig. 2(*c*) for a rather extreme case.

By dividing the area under the relative delay/frequency characteristic into strips of width $2f_1$ (see Fig. 1) it is seen that this area $A_T = 2f_1(T_1 + T_3 + T_5 + \ldots)$, provided only that f_1 is chosen sufficiently low. Thus the peak distortion, expressed as a fraction of the peak-to-peak input, for low values of f_1 is given by the simple relation

$$\Delta V_m = 2A_T \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (5)$$

It is of some theoretical interest to note that the distortion ΔV_m is an integral function [actually $2\int_0^{f_m}(T - T_c)df$] of the phase-delay T, and is not directly related to the group (or envelope or differential) delay $\frac{d\phi}{d\omega} = T + \omega\frac{dT}{d\omega}$.

For the particular case of a rectangular delay characteristic where $T_1 = T_3 = T_5$, etc., we have, for $\phi_n < 36°$,

$$\Delta V = 4T_1(\sin\theta - \sin 3\theta + \sin 5\theta + \ldots)/t_1$$

where $\theta = \pi(2t - T_1)/t_1$. It may be shown (see Section 6) that this expression represents a series of alternating pulses, each having a form given by $\frac{1}{t}\sin\omega_m t$. Combining this result with the equation for amplitude found above, gives [see Fig. 3(*a*)]

$$\Delta V = 2T_1 f_m\left(\frac{\sin\omega_m t}{\omega_m t}\right)$$

for each pulse, where f_m is the highest frequency delayed. Thus the main distortion pulse starts at a time $1/2f_m$ before the transi-

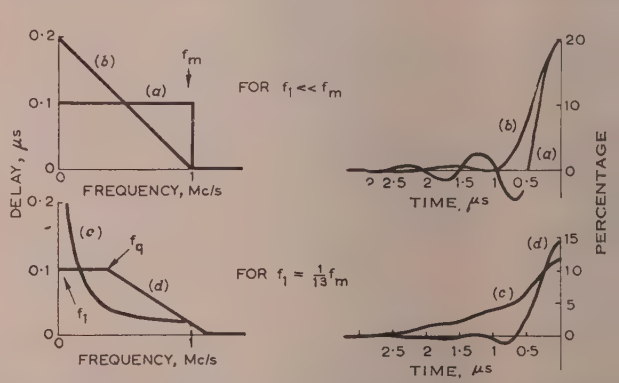


Fig. 3.—Form and amplitude, relative to peak-to-peak of input square wave, of distortion [pre-shoot as in Fig. 2(*b*)] for four examples of relative-delay/frequency characteristic.
Halving T_n or f_m would halve the pre-shoot amplitude.
Halving f_m also doubles the length of pre-shoot.

tion, and has an amplitude $2T_1f_m$. Note also that the area of the pulse is directly proportional to T_1, since its length is clearly proportional to $1/f_m$. The form and amplitude become somewhat different when f_1 is not very much less than f_m, and in the limit, where $f_1 = f_m$, ΔV becomes a sine wave of amplitude $4T_1f_m$.

Another case of interest is the triangular delay characteristic; for $f_1 \ll f_m$ the pulse form is given by

$$\Delta V = T_1 f_m\left(\frac{1 - \cos\omega_m t}{0{\cdot}5\omega_m^2 t^2}\right)$$

as shown in Fig. 3(*b*). Comparing this with the result for a rectangular characteristic having the same area and value of f_m shows that ΔV_m is not altered; the main pulse is rather longer, but the oscillation has virtually disappeared, owing to the absence of any abrupt change in T_n with f.

For the case of constant phase-shift, ϕ, for all harmonics up to the nth, giving $T_n = T_1/n$, for any value of ϕ,

$$\Delta V = \frac{4}{\pi}\sin\frac{\phi}{2}\left[\sin(\omega_1 t - \phi/2) - \frac{1}{3}\sin(3\omega_1 t - \phi/2) + \ldots\right]$$

$$= \frac{4}{\pi}\sin\frac{\phi}{2}\left[\cos\frac{\phi}{2}\left(\sin\omega_1 t - \frac{1}{3}\sin 3\omega_1 t + \ldots\right) - \sin\frac{\phi}{2}\left(\cos\omega_1 t - \frac{1}{3}\cos 3\omega_1 t + \ldots\right)\right]$$

The second term merely shows that the amplitude of the input wave is slightly altered, but the first represents a distortion wave (not a square wave) peaking at $t = t_1/4$ where its amplitude is given by

$$\Delta V_m = \frac{2}{\pi}\sin\phi\left(1 + \frac{1}{3} + \frac{1}{5} + \frac{1}{7} + \ldots\right)$$

Thus, for $\phi < 36°$, as before, the fractional distortion varies from $2\phi/\pi = 4T_1 f_m$ when the input frequency is only just below the highest frequency, f_m, to be delayed (i.e. $n = 1$), up to $4\phi/\pi$ for $n = 7$, thereafter rising gradually with n to approach infinity as f_1 tends to zero. The form of the distortion pulse for this case has been found by plotting the first seven terms, and is as given in Fig. 3(*c*). Another typical case (trapezoidal) has been plotted similarly as in Fig. 3(*d*); this result could have been obtained from the formula given in Section 6 (for f_1 small).

The above analysis gives the distortion ΔV_m at the transition ($t = t_1/4$) for any delay characteristic which can be made to conform with the restriction $\phi_n < 36°$ by a suitable choice of T_c; this limit corresponds to $T_c < 0{\cdot}2$ microsec at 500 kc/s and is not usually exceeded in practical vestigial-sideband systems. [It can be easily shown that the formulae hold (i.e. $\phi < 36°$) for fractional distortions up to 20% for rectangular characteristic, or up to over 20% for the other three types considered.] The analysis is applicable to any shape of characteristic in the low–mid-frequency region, but it cannot be used to find the distortion arising from the delay errors due to high-frequency cut-off, since T_c cannot then be chosen in such a way that the limit $\phi < 36°$ is complied with. Cases where the relative delay is negative (giving pre-undershoot and overshoot) are covered, but if the relative delay were positive in some regions and negative in others, ΔV_m would represent only the distortion at $t = t_1/4$ and not necessarily the peak distortion; in this case the form of the distortion pulse and the position of its peak could differ considerably from that shown in Figs. 2 and 3.

It may be noted that, instead of evaluating the distortion due to extra delay at the lower frequencies, we can evaluate that due to the phase advance ΔT in the high-frequency region in Fig. 1. Similar expressions are obtained, but they are limited as before to $\phi < 36°$ and this limit is, of course, much less easily complied with at the higher frequencies. Moreover, the distinction from the distortion due to cut-off becomes indeterminate, and this method of analysis appears to be of much less quantitative use than that adopted above. Qualitatively, however, it is easier to visualize a short pulse before the transition as arising from the phase advance of a group of relatively high-frequency components rather than from the phase delay of the lower frequencies.

(3) APPLICATION TO TELEVISION

Apart from References 1 and 2 there appears to be little published information on the degree of modulation-frequency delay distortion obtained from practical receiver i.f. and r.f. circuit arrangements. Any attempt to reduce the pre-shoot distortion observed on television pictures must proceed via consideration of the low-mid-frequency delay errors arising from different arrangements of tuned circuits, and this question is therefore worth serious investigation. Some degree of delay distortion is inevitable in television, and the amount is determined by the rate of cut-off which is considered necessary both in the receiver and in the transmitter to avoid interference from (and with) the adjacent sound channel 1·5 Mc/s away from the wanted vision carrier.

It should be noted that this delay distortion is, in practice, more serious than other types of pulse distortion arising in the r.f. or i.f. circuits of a receiver, in that the simpler video correction circuits are totally ineffective against it. A bridged-T phase corrector is a satisfactory remedy, but there is usually no convenient point at which it can be inserted in a conventional broadcast receiver, with the result that an appreciable increase in cost would be involved.

In the application of the theoretical results to practical television pulse work, the effect of the other sources of distortion must be kept in mind, namely

(*a*) The observed pre-shoot and smear should agree substantially with that calculated from a measured system video delay characteristic when the region of low-mid-frequency delay distortion is quite distinct from that of cut-off distortion, as in Fig. 1. As the cut-off frequency (conveniently specified by the 3 dB point, f_c, on a video modulation test) approaches f_m the visible smear will tend to be obscured, although the distortion is, of course, still present. Some indication of this effect may be obtained from Fig. 2(*c*) in which the slope of the transition is drawn to correspond to a value of $f_c = 2$ Mc/s and the delay characteristic is triangular with $f_m = 2$ Mc/s and $T_1 = 0{\cdot}1$ microsec.

(*b*) When the fractional modulation is high (say over 50%) the effects of quadrature or non-linear distortion appear, being visible chiefly as a tendency to increased overshoot on the rise and pre-shoot on the fall of the i.f. envelope, plus a rounding of the start of the rise and the end of the fall. Fortunately, however, the black/white modulation ratio involved in actual pictures is not sufficient to make these effects noticeable, except where the modulation falls below, say, 15% at peak white on a negative modulation system.

(4) ACKNOWLEDGMENT

The author desires to thank Mr. L. Barclay for his help in calculating two of the curves, and Messrs. E. K. Cole, Ltd., for permission to publish the results of work done in their laboratories.

(5) REFERENCES

(1) Kell, R. D., and Fredendall, G. L.: 'Standardisation of the Transient Response of Television Transmitters', *RCA Review*, 1949, **10**, p. 17.

(2) Avins, J., Harris, B., and Horvath, J.: 'Improving the Transient Response of Television Receivers', *Proceedings of the Institution of Radio Engineers*, 1954, **4**, p. 274.

(3) Cherry, C.: 'Pulses and Transients in Communication Circuits' (Chapman and Hall, London, 1949).*

(6) APPENDIX

The spectrum for any series of pulses with spacing t_1 may be written

$$V_1 = a_1\cos\omega_1 t + a_2\cos 2\omega_1 t + a_3\cos 3\omega_1 t + \ldots$$

A similar series of pulses, but shifted in time by $t_1/2 = \pi/\omega_1$ and inverted, is given by

$$V_2 = -\left[a_1\cos(\omega_1 t + \pi) + a_2\cos 2(\omega_1 t + \pi) + \ldots\right] = a_1\cos\omega_1 t - a_2\cos 2\omega_1 t + a_3\cos 3\omega_1 t + \ldots$$

Thus the sum of the two, giving the required alternating series of pulses [as, for example, in Fig. 2(*a*)] spaced by $t_1/2$ is

$$V_1 + V_2 = 2(a_1\cos\omega_1 t + a_3\cos 3\omega_1 t + a_5\cos 5\omega_1 t + \ldots)$$

* Dr. Cherry gives a general treatment of the relations between steady-state and pulse characteristics and extends this treatment to cover vestigial-sideband systems. He does not refer specifically to the delay distortion primarily considered in the paper.

Shifting each pulse in time by $3t_1/4 = 3\pi/2\omega$, gives

$$V_3 = 2[a_1 \cos(\omega_1 t + 3\pi/2) + a_3 \cos 3(\omega_1 t + 3\pi/2) + \ldots]$$
$$= 2(a_1 \sin \omega_1 t - a_3 \sin 3\omega_1 t + a_5 \sin 5\omega_1 t + \ldots)$$

If $a_1 = a_2 = a_3$, etc., terms up to $n\omega_1 t = \omega_m t$ are included and n is very large,

$$V_1 = \int_0^{\omega_m} \cos \omega t d\omega = \frac{1}{t} \sin \omega_m t$$

and this gives the waveform of each pulse for a rectangular spectrum up to ω_m.

For a triangular spectrum the waveform is

$$V_4 = \int_0^{\omega_m} (1 - \omega/\omega_m) \cos \omega t d\omega = \frac{1 - \cos \omega_m t}{\omega_m t^2}$$

and for a trapezoidal spectrum it is

$$V_5 = \int_0^{\omega_q} \cos \omega t d\omega + \int_{\omega_q}^{\omega_m} (1 - \omega/\omega_m) \cos \omega t d\omega$$
$$= \frac{\omega_q}{\omega_m} \frac{1}{t} \sin \omega_q t - \frac{1}{\omega_m t^2} (\cos \omega_m t - \cos \omega_q t)$$

Here ω_q is the upper limit of uniform delay, as shown in Fig. 3.

621.374.32: 621.389.620.113

Paper No. 2053 M
July 1956

AN ELECTRONIC MACHINE FOR STATISTICAL PARTICLE ANALYSIS

By H. N. COATES, Ph.D., B.Sc.(Eng.).

(*The paper was first received* 1*st December*, 1955, *and in revised form* 23*rd January*, 1956.)

SUMMARY

A system is described for associating and collecting the intercepts of individual particles in a particle scanning system, where the information is presented as a function of the scanning voltages. A series of stores is used to segregate the intercepts, each store having its own memory system and provision for re-use on completion of the scanning of the particle with which it is associated; the stores can thus be used many times during a single frame scan. A method of adding the intercepts of each particle to obtain a measure of the area of the particle is described, but this must be regarded as only one of the possibilities of extracting information from the series of intercepts collected.

(1) INTRODUCTION

The line-by-line scanning apparatus described by Roberts and Young[1] produces information about a scanned particle which enables the particle shape to be completely reconstructed on a cathode-ray-tube screen. The information is in the form of a signal with two states representing, respectively, the spot on a particle and not on a particle. The particle intercepts contain all the necessary information about particle shape and size, and the difficulty has been to extract from this series of intercepts details of that particular feature of the particle which it is desired to explore.

A particular particle intercept must first be recognized and its association with other intercepts on previous lines established. The general approach to this problem has been to provide a memory system which stores the line information,[2] or to read two lines simultaneously and thus enable a correlation to be achieved. Theoretically no complete line-to-line memory device is necessary, as the beginning of a particle intercept is associated in time with the beginning of the previous intercept, and if a time memory is provided a further memory system is not necessary. This has the advantage of obviating a split-beam scanning device with its associated disadvantages, or a line-to-line cyclic memory device which is invariably fixed in its time cycle and does not allow flexibility of the scanning system to cope with differing-density samples of particles and different degrees of resolution required.

The system outlined here is based on a memory device which stores the line-scan voltage associated with the commencement of a particle intercept and provides warning when this voltage is approached on a subsequent line, in order to enable the next particle intercept to be directed into the same store as the previous one. In this manner it is possible to collect the intercepts associated with any one particle in a storage device, and this can be done for a number of particles equal to the number of stores provided.

The particular aspect considered here is the summation of the intercepts associated with any one particle to provide a measure of the area of the particle to the limitations of scanning-spot size and line spacing. To achieve this, the particle intercepts are added in each individual store and the final sum total is discharged into a classifying unit on completion of particle scan. The sum total of the particle intercepts appears as a voltage which is classified by the unit in terms of preset levels to provide an analysis of the number of particles in any frame lying between predetermined areas. As the discharges to the classifying unit occur at the completion of each individual particle scanned, the information in this unit expressed in terms of the frame scan voltage gives an indication of the size distribution of particles throughout the sample. The number of discharges gives a direct count of the number of particles scanned.

(2) PRINCIPLES OF OPERATION

Fig. 1 shows a block schematic of the complete system. Signals from the photocell are fed into a command unit where they are amplified and chopped. This unit also generates a line-scan voltage and a frame-scan voltage, and these, together with pulses derived from the beginning and end of the line flyback, are fed via bus-wires to all the stores in a batch associated with the particular equipment. The stores are also connected to a common line terminating in the classifying unit.

A gate circuit is fitted at the entry to each store (see Fig. 2) and operated from the store's memory circuit. The intercept information, consisting of a negative-going 'pip' corresponding to the commencement of a particle scan, is fed to the first store, which will accept it if its gate is open, or pass it on to the next store if the gate is closed. In this way the intercept information travels down the line of stores until it finds an open gate which corresponds to the store for which it was intended, or one which is idling and ready to accept fresh information as the case may be.

When information is absorbed through an open gate nothing is passed on to the next store, and the gate closes immediately so that future information will continue down the line of stores until the correct one is reached. At this instant information is sent to the memory circuit in the store, which closes and holds the line-scan voltage associated with the commencement of the intercept information, and the store integrator is started. At the end of the intercept the integrator is stopped by a pulse sent down No. 3 bus-wire. No guidance of this pulse is required, as one integrator only is operating at a time, and the pulse consequently appears at all integrator switches.

In the next line, the store memory will open a predetermined interval before the value of line-scan voltage at which the beginning of the intercept information was received in the previous line, and this will open the gate at the entry to the store ready to absorb the next piece of information. At the same time a pulse is sent down No. 2 bus-wire to close any gates which may be open in preceding stores, so that the information will be handed on until this store is reached.

The cycle of events is then repeated, the memory storing the new value of line-scan voltage associated with the commencement of the new particle intercept. In this way the profile of the explored particle is followed with a constant anticipation tolerance, defined as the difference between the value of time or voltage associated with the commencement of a particle intercept and that value on the succeeding line when warning is given of the pending arrival of further information associated with the same particle (the time or voltage being measured along the lines in each case).

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.

Dr. Coates was formerly with E.P.S., Ltd., and is now at the Mullard Research Laboratories.

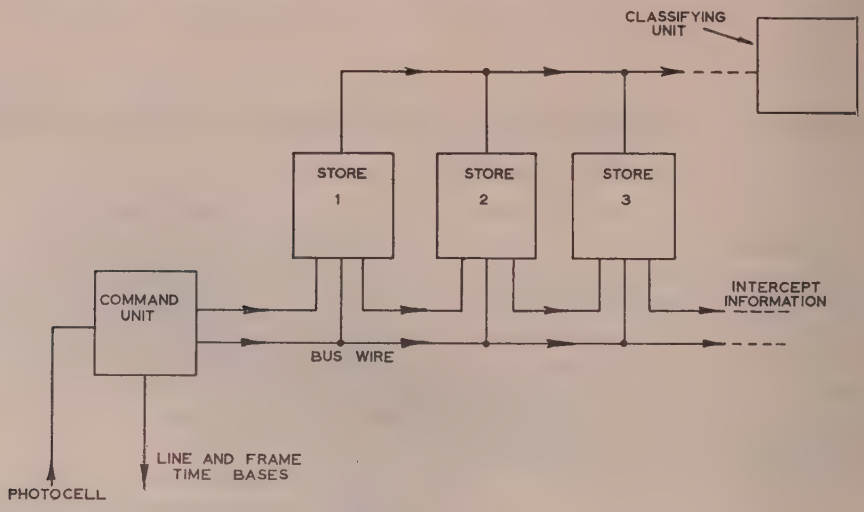


Fig. 1.—Block schematic of the system.

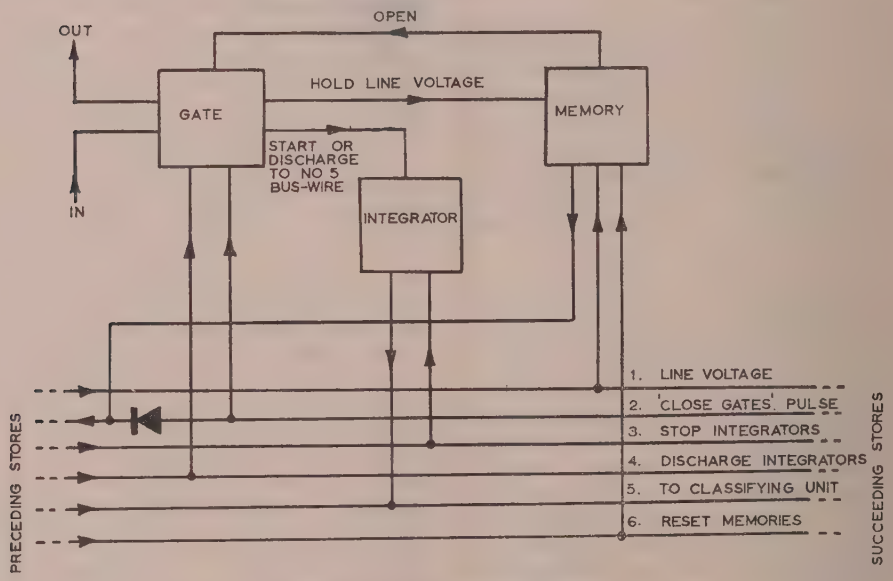


Fig. 2.—Block schematic of a store.

Each store is arranged so that it will continue operating until all the intercepts associated with the particle have been collected. When the store opens on the next line and no further particle information is received within a given time, the store discharges the integrator into No. 5 bus-wire and returns to an idling condition ready to accept fresh intercepts. This means that the stores are used over and over again, and, with some exceptions, the number of stores required is equal to the number of particles to be scanned in any one line.

When a store is idling its memory is reset via No. 6 bus-wire at the commencement of each new line. Should No. 2 store be operating and become due to receive the next piece of intercept information while No. 1 store is idling, it will arrange to close No. 1 store gate as described earlier, and No. 1 store will then be out of commission for the remainder of the line until it is reset. Thus a condition can arise when a store towards the end of the batch is collecting information about a particle at the beginning of the line scan and closing all idling stores in front of it for most of the line duration; this state of affairs will continue until the store has completed its integration. This means that the number of stores available for receiving fresh information is limited to the idling stores beyond this one in the batch which can give rise to conditions where a greater number of stores is required than particles in any line scan.

Stores idling further down the batch than a store operating are not affected in any way by the operation of the store. No. 2 bus-wire contains a unidirectional device in each store to ensure that 'close gate' pulses are passed forward only, leaving idling stores beyond an operating store free to collect new intercept information.

(3) DESIGN OF THE STORES

An important feature of the design of the stores is that there is no cyclic function generated or required in connection with these, and their operation is consequently largely independent of the line or frame frequency of the scanning system, thus allowing

wide flexibility in use. For this reason the line-scan voltage is used as a reference throughout the discussion of the system; and while this can conveniently be regarded as a linear function of time, it need not necessarily be the case, as non-linearity does not affect the principles of design.

(3.1) Gating Circuit

Fig. 3 shows the essential features of the gating circuit. V3 and V4 form a bistable circuit which switches grid 3 of V2 and grid 1

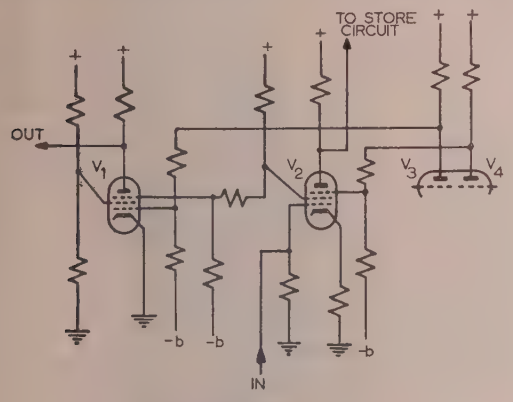


Fig. 3.—Gating circuit.

of V1. V1 and V2 are short-suppressor-grid-base valves, intercept information being fed to the control grid of V2 from the preceding store, and when the store is in operation being passed on to the succeeding store via the anode of V1.

When the store is idling, V4 is non-conducting and the bias is removed from grid 3 of V2. V3 is conducting heavily and V1 is cut off on grid 1. Negative pulses fed to grid 1 of V2 are passed into the store from the anode. As V1 is cut off no pulses are passed on to the succeeding stores.

When the store is operating, V3 is cut off and consequently the bias is removed from grid 1 of V1. However, V4 is conducting hard and V2 is cut off on grid 3 and is passing a large screen current, with the result that the screen potential is low and V1 is arranged to be cut off on grid 3 under these conditions. A negative pulse applied to V2 of sufficient amplitude to cut the valve off momentarily does not appear at the anode and consequently does not affect the store in any way. The pulse produces a positive pulse on the screen grid of V2 which removes the bias from grid 3 of V1, producing a negative pulse at the anode which is passed on to succeeding stores.

The system is arranged to have a small pulse gain from grid 1 of V2 to the anode of V1, so that the pulse amplitude travelling down the line of stores is always limited by the grid base of V2 and does not appreciably change in amplitude. Another feature of the system is that opening or closing the gate V2 does not produce a transient which could be passed down the line of stores and upset their operation.

(3.2) The Integrator

Fig. 4 is a diagrammatic representation of the store integrator circuit, with V3 as one half of a bistable circuit. If C be fully charged to the h.t. potential, then, when V3 conducts, C will discharge through V1 and R1, and, as the resistance of R2 and the slope resistance of V3 are small compared with that of R1, the discharge time-constant is approximately CR_1. The discharge will continue while V3 is conducting, and terminate when the anode of V3 is returned to the h.t. potential. Each time the bistable circuit operates to make V3 conduct, C will discharge in this manner towards the conducting anode voltage of V3.

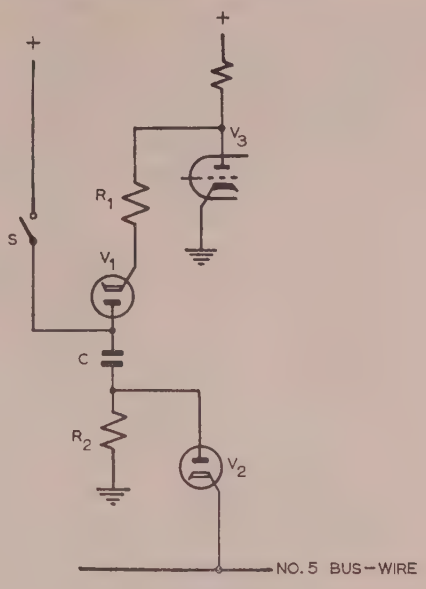


Fig. 4.—Integrator circuit.

These discharges represent the intercepts of a scanned particle; the duration of each discharge being proportional to the length of particle intercept. The resistance of R1 is arranged so that only a reasonable proportion (say 50%) of the charge on C is removed.

On the completion of the integration the switch S is closed, and current flows into C through R2 to restore the voltage across C to h.t. potential. This produces a small voltage across R2, the amplitude of which is proportional to the current flowing into C, which, in turn, is proportional to the charge removed during the integrating process. This pulse is fed into No. 5 bus-wire and thence to the classifying unit. The amplitude of this pulse will not be directly proportional to the total time of particle scan, but will always be connected to it by a common law which can be corrected for by a weighting network between No. 5 bus-wire and the classifying unit, or alternatively by modifying the preset levels in the latter.

(3.3) The Memory Circuit

Fig. 5 is a diagrammatic representation of the memory circuit. The triode V is supplied on its anode with a waveform, corre-

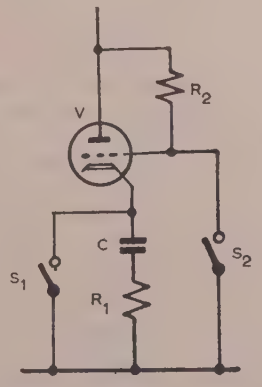


Fig. 5.—Memory circuit.

sponding to the line time-base, having an amplitude starting at approximately 10 volts and rising to approximately 100 volts. During the rise time of the voltage the valve passes a current to charge C which follows the line voltage less the drop across the valve and across R1. At any instant during the line scan the closing of S2 will impose a negative bias on the valve sufficient to prevent further current passing, and the cathode potential will consequently 'freeze' at the line voltage at the instant the switch occurs, and as the memory circuit resets itself every line whether the store is idling or operating, slow changes in line-scan amplitude are automatically compensated for and do not upset the accuracy of the system.

(4) OPERATION OF THE STORES

Fig. 6 is a schematic of the complete store, and Fig. 7 shows some of the waveforms associated with various parts of the circuit.

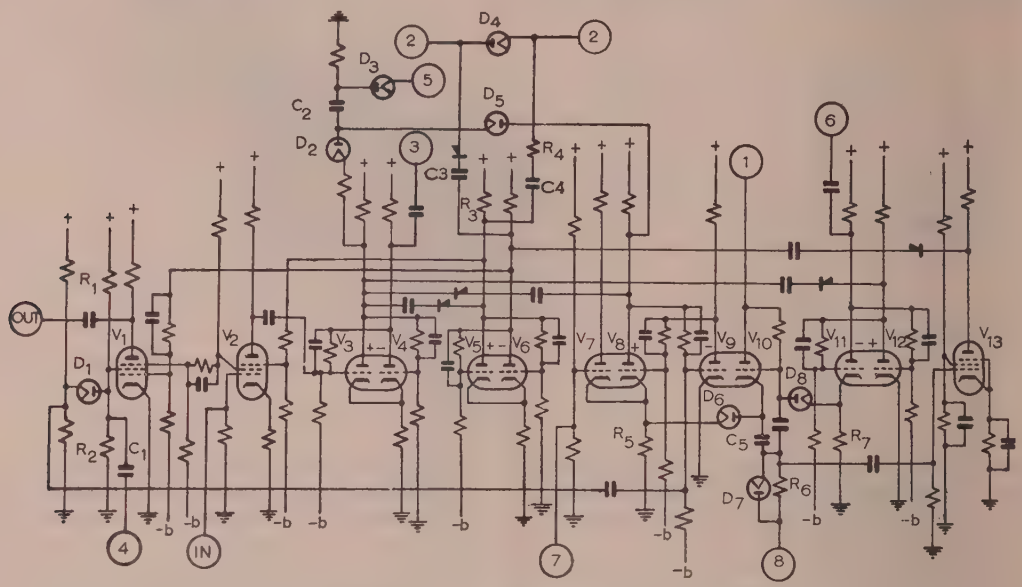


Fig. 6.—Circuit diagram of a store.

was closed less the drop in R1 and across the valve (in practice the latter is small).

S2 is arranged to open at the beginning of the next line scan, but no current will pass through the valve until the anode and grid potentials approach the cathode potential fixed by the voltage across C, when current will again flow into the capacitor. This causes a pulse across R1, which signifies that the memory circuit has opened, and produces appropriate switching in the store.

If the store is idling and the memory is not operated during the line scan, S1 is arranged to close at the beginning of the line flyback and remain closed for the duration of the flyback, during which time the capacitor has discharged ready to commence recharging in the next line. The capacitor C will discharge at its own natural time-constant shunted by the heater-cathode resistance of the valve (which can be arranged to be very high), but this time-constant can be large for a suitable type of capacitor, causing negligible decay between lines.

In practice, further complications are added to the memory circuit to ensure satisfactory operation. The pulse across R1 is fed through a small capacitor to the control grid of the valve, so that the instant at which the valve starts to conduct is clearly defined along the line scan. A diode is connected across R1 to remove the large negative pulse when the capacitor is discharged during the line flyback, as this tends to saturate the following amplifier. The bottom end of R1 is taken to No. 8 bus-wire, which is fed with the frame time-base in a negative-going sense. This slowly reduces the charge on the capacitor and ensures that the memory circuit opens before the voltage at the cathode of the valve reaches that at which it closed in the previous line.

Both line and frame time-bases are fed into the bus-wires from cathode-followers to ensure that no interaction between stores

Apart from the circuits already described the store consists for the main part of bistable circuits[3] of two types—common cathode and separate cathodes. V3/V4 and V5/V6 represent examples of the former type and V8/V9 and V11/V12 are examples of the latter type. V10 is the memory valve and C5 the memory condenser, as already described. The potentials marked at the anodes of the bistable valves represent the quiescent condition when the store is empty and idling. Under these conditions V10 is passing the line-time-base waveform [Fig. 7(*a*)] and charging C5. V7 is conducting heavily and producing a large bias across R5, thus ensuring that D6 is cut off for the complete cycle of the waveform at the cathode of V10. A negative pulse is applied to the grid of V7 for the duration of the line flyback from No. 7 bus-wire, cutting off V7. As V8 is the cut-off section of a bistable circuit, no current flows through R5 during the line flyback, allowing C5 to discharge through D6 practically to earth potential. The initial rise of the line time-base before commencement of line scan ensures that any residual voltage remaining on C5 is insignificant compared with the starting value of the line voltage.

Should V8/V9 change its state of stability so that V8 is conducting heavily, sufficient voltage will be produced across R5 to prevent C5 discharging, irrespective of the current through V7. This ensures that, when the store is operating, the memory circuit is not discharged during the line flyback. The state of stability of V8/V9 indicates whether or not the store is operating.

Let it be assumed that the store is empty and idling and the potentials of the anodes of the bistable circuits are as indicated. A negative pulse is received on the grid of V2, which represents the commencement of a new particle scan [Fig. 7(*c*)]. This appears as a positive pulse at the grid of V3, causing a change of

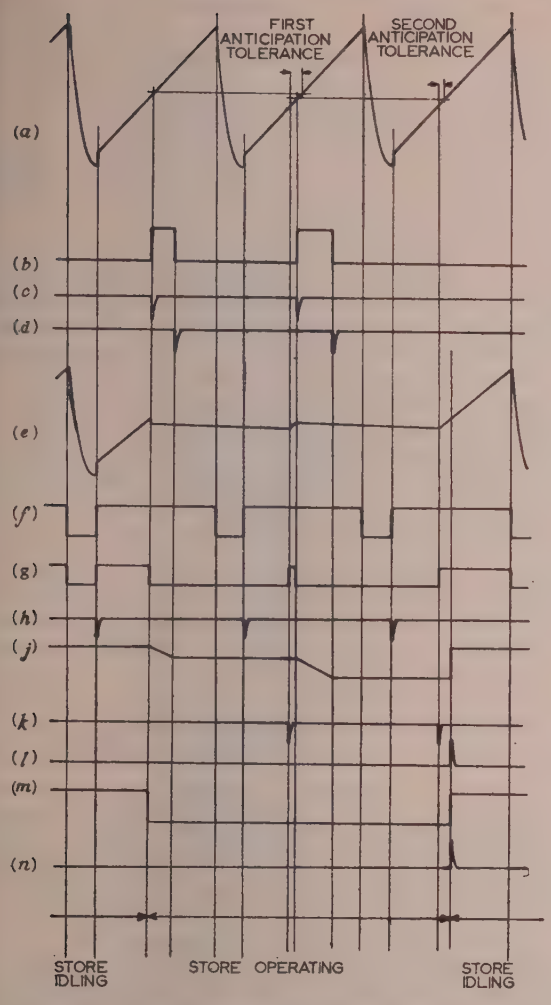


Fig. 7.—Typical waveforms associated with the stores.

(*a*) No. 1 bus-wire waveform.
(*b*) Chopped output from photocell.
(*c*) Intercept information.
(*d*) No. 3 bus-wire waveform.
(*e*) Waveform at cathode of V10.
(*f*) No. 7 bus-wire waveform.
(*g*) Voltage across R6.
(*h*) No. 6 bus-wire waveform.
(*j*) Waveform at anode of D2.
(*k*) No. 2 bus-wire waveform.
(*l*) No. 4 bus-wire waveform.
(*m*) Voltage at anode of V8.
(*n*) No. 5 bus-wire waveform.

state of this circuit. The anode voltage of V3 goes negative taking with it the anode voltage of V5, and thus causing change of state of V5/V6. The anode voltage of V5 goes negative, cutting off the gate valve as already described.

The negative pulse on the anode of V3 also changes the state of V8/V9 and V11/V12. The former circuit ensures that the memory will not be discharged, and the latter circuit now cuts off V11, which returns its cathode to earth potential and thus 'clamps' the grid of V10 to earth via D8, stopping V10 from conducting and 'clamping' the cathode potential. This potential will drop by an amount equal to the *IR* drop across R6, which is generally of the order of $\frac{1}{2}$ volt [see Fig. 7(*e*)]. This is important as it allows an additional anticipation tolerance between the first intercept scan and the second, where the maximum anticipation is normally required.

While V8 was cut off (the idling condition) C2 was maintained charged to h.t. potential via D5. On commencement of operation, the potential of the anode of V8 drops for the total duration of intercept summation, and C2 is allowed to discharge through D2 whenever the anode potential of V3 goes negative. The store continues with C2 discharging until the integrator is stopped by a pulse down No. 3 bus-wire common to the anode circuits of all V4 valves in the batch of stores.

At the commencement of the next line scan, No. 6 bus-wire transmits a pulse to the anode of V11, resetting this bistable circuit so that a voltage is developed across R7 which is of greater amplitude than the line scan, allowing the grid of V10 to follow the anode during the next line. C5 will have discharged slightly owing to the frame scan on No. 8 bus-wire, and V10 will consequently start conducting a little before the voltage at which it was previously clamped. This causes a pulse across R6 [Fig. 7(*g*)], which is amplified in V13 and operates the bistable circuit V5/V6. This arranges for the gate to be opened, and the store now awaits the next pulse at the grid of V2. During this waiting period C5 is being charged from the line time-base, so that as soon as the pulse arrives it is clamped at the new line potential associated with the commencement of this intercept. The operation of V5/V6, causing the anode potential of V6 to go negative, sends a negative-going pulse via C3 to preceding stores along No. 2 bus-wire. This enters preceding stores via their capacitors C4, and will thus set V5/V6 in these stores so that the gate valve V2 is closed if the store should happen to be idling. This ensures that the wanted piece of information is passed down the line of stores until it reaches the one whose memory has just operated. D4 prevents pulses being passed to succeeding stores in the batch.

The pulse travelling down No. 2 bus-wire will continue until it ultimately reaches the command unit. This is an indication that a store memory has operated and the store is awaiting information. If such information is not forthcoming within a predetermined time (twice the anticipation tolerance), it is an indication that the summation of intercepts has concluded for that particular particle. The command unit sends a positive-going pulse down No. 4 bus-wire after this predetermined interval unless intercept information has previously been sent out. The store which has completed its integration is waiting with V2 open and V1 cut off on grid 1. Only stores with their gates open (i.e. idling empty stores beyond the one in question) also have V1 cut off on grid 1. These stores have the screen grid of V1 at a relatively high potential fixed by R_1 and R_2, and under these conditions D1 is arranged to be on the point of conducting. Consequently a positive pulse applied through C1 appears at the cathode of D1, and is transmitted to the bistable circuit V8/V9 causing a change of state. This change of state will only occur in a store which has been operating—consequently only one store will be affected by the pulse sent along No. 4 bus-wire, i.e. the one which has just completed its integration. Stores with their gates closed have the screen grid of V1 conducting, and its potential under these conditions is arranged to be too low to allow D1 to conduct when the pulse is applied.

The change of state of V8/V9 returns the anode of V8 to h.t. potential, causing a charging current to flow through D5 and the pulse for classification to be sent out by D3 as already described. The store is now empty and idling again with its gate open and its memory circuit following the line time-base awaiting fresh particle information.

Idling stores which have been closed during a line scan by succeeding stores are opened again at the beginning of the next line when V10 starts conducting and a positive pulse appears across R6. This operates V5/V6, sending a negative pulse forward down No. 2 bus-wire from the anode of V6, and a positive pulse to succeeding stores down the same bus-wire from the

anode of V5. This positive pulse will be cancelled very largely by the negative pulse from opening succeeding stores, resulting in little or no change in the bus-wire potential. This ensures that preceding stores are not closed by the negative pulse from V6, and they in turn will be generating their positive pulse at the same instant.

No. 2 bus-wire is arranged to terminate in a high impedance at the command unit end and a relatively low impedance at the other end, thus ensuring that during the line scan an opening store will not send a positive pulse down No. 2 bus-wire to succeeding stores interfering with their gating operation. Under these circumstances the positive pulse from the anode of V5 is dissipated in R4, which is arranged to have a resistance high compared with the low terminating impedance of No. 2 bus-wire and low compared with that of R3, so as not unduly to attenuate negative pulses entering the store to close the gate during normal operation.

(5) RESOLUTION

Fig. 8 represents a particle being scanned in the conventional manner (top to bottom, left to right). The left-hand side of the

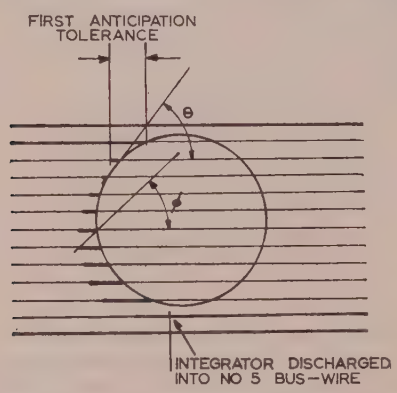


Fig. 8.—Particle being scanned by the system.
The heavy lines indicate the periods for which the store gate is open.

particle profile, as recognized by the apparatus, is a line joining the commencements of adjacent particle intercepts, and this line subtends an angle θ to the direction of line scan.

When the particle is scanned linearly with a constant anticipation tolerance, the line joining the points at which the store gate opens forms a 'shadow' of the particle profile projected at an angle ϕ to the direction of line scan. This angle, known as the 'angle of anticipation', is the minimum value of θ that can be followed by the system subsequent to the second line intercept. It is the angle ϕ which should be maintained constant and not the actual anticipation tolerance, which will need to vary with line spacing and particle size.

By definition, $\tan \phi$ = (line spacing)/(anticipation tolerance), and both these quantities are proportional to the progression of frame-scan voltage between lines. Assuming the anticipation tolerance to be small compared with the total line-scan voltage,

$$\phi = \text{arc tan (line-deflection sensitivity)/(frame-deflection sensitivity).}$$
$$= \text{arc tan } (Bx/Ay)$$

where B/A = (Frame amplitude)/(line amplitude) = Aspect ratio of scanned sample.

x = Line voltage amplitude
y = Frame voltage amplitude
} fed to stores.

For example, for a 90-volt line amplitude and an aspect ratio of unity, over 500 volts frame scan is required for a 10° angle to the line scan subtended by the scanned-particle profile, subsequent to the second scanning line.

This angle is independent of line or frame linearity of scan or the ratio of line to frame velocity, but this ratio will affect the number of intercepts associated with a given particle and consequently the overall accuracy of the system.

The line-voltage amplitude should be sufficient to allow a reasonable change during a particle intercept, and suggested figures are 3 volts per particle, providing storage for 30 particles per line in the case referred to above. Using these figures, a line spacing giving an anticipation tolerance in the region of $\frac{1}{4}$ volt is reasonable, which means that the memory system must open to considerably closer tolerances than this if the anticipation tolerance is to be at all constant and reliable. An accuracy of $\pm 0{\cdot}05$ volt was obtainable repeatedly with the memory system described, without much difficulty, and should some information be available about the approximate range of sizes of particles in a sample to be scanned, it is quite possible that a figure of 1 or 2 volts per particle could be used. This means that a smaller frame-voltage amplitude would be necessary; alternatively, an aspect ratio of 2 : 1 or 3 : 1 could be used and something like 2000 particles assessed at one frame scan.

At a line frequency of 200 c/s the anticipation tolerance represents 12 microsec (assuming linear scanning voltage) and the current through the memory condenser C5, as given by the rate of change of charge, is approximately 0·04 mA when C_5 is $0{\cdot}002\,\mu$F. This current must be supplied from No. 1 bus-wire, and with R_6 equal to 10 kilohms it produces an extra 0·4 volt anticipation tolerance on the second line intercept. For this to be constant, C5 will require changing for different line-scan rates, but it is the only component in the store to be so affected.

(6) CONCLUSIONS

The system described is capable of handling any number of particles, limited only by the size of the equipment and the available amplitude of the frame voltage. This latter is of particular importance as it allows more economical use to be made of a given number of stores.

The system has not produced a satisfactory answer to the difficulties present when re-entrant particles are scanned, and under these circumstances a particle may be counted more than once and false representations of its area given. The system is considered to have its best application where statistical analysis is required of large numbers of particles whose shape is predominantly circular or elliptical, as might result from surface tension or similar effects.

Although such a system as has been described is necessarily elaborate and can be justified only when precise information about the particles is required, it does enable an almost unlimited amount of such information to be extracted from the particle intercepts. Once the store has been constructed, such devices as the integrator or arrangements for recording maximum particle dimensions, etc., are relatively simple additions, and an approach is possible towards the more elaborate measurements of particle shape so long awaited in both the medical and the industrial worlds.

(7) REFERENCES

(1) Roberts, F., and Young, J. Z.: 'The Flying-Spot Microscope', *Proceedings I.E.E.*, Paper No. 1348 R, April, 1952 (**99**, Part IIIA, p. 747).

(2) Dell, H. A.: 'Stages in the Development of an Arrested Scan Type Microscope Particle Counter', *British Journal of Applied Physics*, Supplement No. 3, 1954, p. 156.

(3) Renwick, W., and Phister, M.: 'A Design Method for Direct Coupled Flip-Flops', *Electronic Engineering*, 1955, **27**, p. 246.

621.376.22: 621.318.134

Paper No. 2061 R
July 1956

A FERRITE MICROWAVE MODULATOR EMPLOYING FEEDBACK

By W. W. H. CLARKE, Ph.D., B.Sc., Associate Member, W. M. SEARLE, M.Sc., and F. T. VAIL, B.Sc.

(*The paper was first received* 18*th October*, 1955, *and in revised form* 26*th January*, 1956.)

SUMMARY

The amplitude modulation of microwaves produced by the magnetization of ferrites is non-linear and suffers from hysteresis. Hence, square-wave modulation is the only function in which the distortions are not objectionable.

The paper describes a feedback method of applying the modulation signal, providing a linearity substantially that of the feedback crystal used to detect the modulated microwave signal, and reducing the effect of hysteresis by an amount approximating the feedback loop gain. Pure sine-wave modulation is achieved at low frequencies, in which the second-harmonic sidebands are more than 45 dB below the fundamental. Linear modulation, by sawtooth and square waveforms, is also achieved, in which the modulation envelope faithfully reproduces the applied signal.

The employment of ferrite microwave modulators in engineering applications will involve techniques which are already standard for lower frequencies with conventional components, and the paper establishes the feasibility of using the powerful method of envelope feedback to control them.

(1) INTRODUCTION

Amplitude modulation of microwaves has generally been carried out by modulation of the oscillator—a process which has many disadvantages, including severe incidental frequency modulation. To satisfy a requirement for pure amplitude modulation of a microwave signal at audio and low-supersonic frequencies, an investigation of the use of a ferrite microwave modulator with envelope feedback has been carried out.

Useful gyromagnetic effects at microwave frequencies are now well known[1,2,3] and components employing the Faraday and related effects[2–8] have been reported. Such modulators with variable magnetizing fields all respond non-linearly and show hysteresis, and they are thus not directly suitable for applications requiring linear response. Again, most commercial modulators are designed for very low frequencies of modulation.

The paper describes the application of feedback to a ferrite modulator specially developed for high-speed switching,[6] which provides a high degree of linearity over a band of applied frequencies, and the performance is illustrated by a variety of modulation waveforms. It is shown that the design is within the limitations of feedback circuits for application with ferrite modulators, but the bandwidth and feedback gain are noise-limited by the detection crystal. A less noisy detector would alter this fundamental limitation, permitting increases in feedback gain and bandwidth.

(2) THE FERRITE MODULATOR

The amplitude modulator makes use of the Faraday effect in a circular guide interposed between rectangular guides. Fig. 1 is a photograph of the modulator. It has magnetizing coils spaced from the wall of the circular guide by foam plastic to minimize shunt capacitance, which impairs the response of the coil to applied voltage. The ferrite is 0·2 in in diameter and

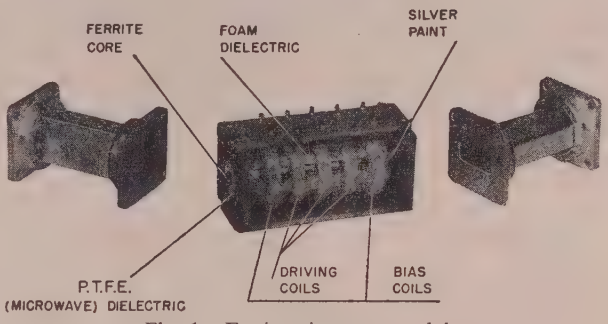


Fig. 1.—Ferrite microwave modulator.

4 in long and is mounted in a 4 in section of solid polytetrafluoroethylene 0·75 in in diameter, with a waveguide wall produced by painting the p.t.f.e. rod with silver paint and forcing waveguide-coupling flanges over the paint at the ends. The type of ferrite used is MF1331, which is reported[9] to be composed of 60–65% MgO, 5–10% MnO and 30% Fe_2O_3. The required rotations are achieved with fields in the neighbourhood of 30 oersteds. Eddy-current screening of the core from the magnetizing field is minimized by keeping the paint layer as thin as possible, consistent with microwave transmission.

When the ferrite is magnetized along the direction of propagation, the Faraday effect is manifested by anti-reciprocal rotation of the plane of polarization of the microwave signal.* Hence, the component polarized in the direction accepted by the output waveguide is made to vary. The rotation is proportional to intensity of magnetization M and path length l, namely the

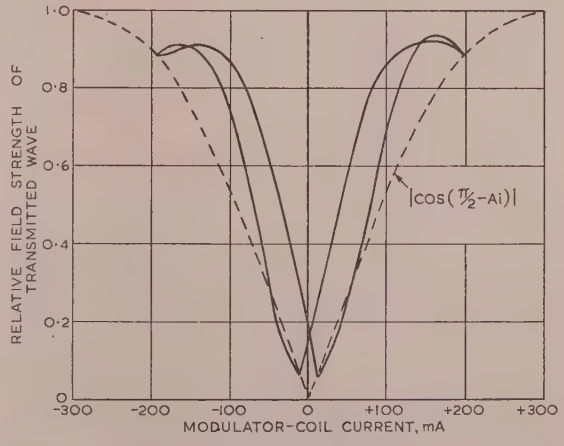


Fig. 2A.—Switch transmission cycle compared with the function

$$\left|\cos\left(\frac{\pi}{2} - Ai\right)\right|.$$

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.

The authors are at the Defence Research Telecommunications Establishment, Ottawa.

* It is conventional to explain the plane-wave rotation by a difference in the velocities for equal components of right-hand and left-hand circularly polarized waves, assumed to be formed when the plane wave enters the circular guide.

length of the ferrite. The transmitted energy is therefore $W_0 \cos^2(\theta - KIM)$, where W_0 is the incident energy, K is a constant and θ is the angle between input and output rectangular guides ($\theta = \pi/2$ in this case).

Hysteresis being neglected, the magnetization is proportional to coil current i, and hence the transmitted microwave field-strength is proportional to $|\cos(\pi/2 - Ai)|$, where A is a constant. Fig. 2A shows experimental curves for the modulator compared with $|\cos(\pi/2 - Ai)|$, for a value of A derived experimentally from the separation of successive transmission minima. The omission from the modulator of absorbing vanes, to remove the wave polarized at 90° to the output guide, permits multiple reflections, resulting in a greater modulation sensitivity. This is due to the fortuitous reinforcement of the transmitted wave by the reflections as soon as a small rotation is introduced; hence

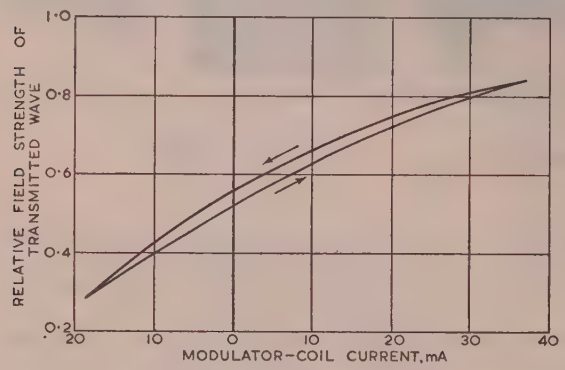


Fig. 2B.—Characteristic modulation cycles.

the deviation in slope of the practical curves from the theoretical curve. This reinforcement in one part of the microwave spectrum implies cancellation effects in others, so that a bandwidth penalty is involved, which, however, is unimportant in the present application. Fig. 2B shows a magnetization cycle typical of those used for modulation.

(3) THE FEEDBACK PROBLEM

A feedback circuit based on linear detection of amplitude modulation has been constructed, and performs with a linearity varying with percentage modulation. Owing to the nature of the transmission characteristic, the required change in magnetization for a given change of output is large near maxima and smaller near minima. The circumstances governing the feedback stability therefore change during a cycle.

It is necessary to distinguish between cases with the same modulation depth but using different proportions of the available microwave energy by operating under different bias conditions. A standard criterion of bias can be adopted for evaluation purposes, whereby the static bias corresponds with a transmission of half the field strength at maximum. From the half-field criterion bias, modulation depth approaching 100% and utilizing the maximum energy can be produced without change of bias in the presence of a suitable feedback; but for practical purposes this is not the optimum of power utilization. The bias actually used is a suitable compromise between that required for optimum linearity and optimum power utilization.

The approach to zero transmission is particularly critical, since a very small overshoot of magnetization can lead to loss of control by the amplifier, while a similar, less critical overshoot can occur at the maximum of transmission. These factors are apparent in Fig. 2A, and underline the importance of operating-point stability. Though it is not beyond present techniques to produce an operating-point feedback with a d.c. amplifier to control the bias, the additional complication is undesirable, and it is better to leave some margin at upper and lower transmission levels. Such margins are better for linear modulation in any case, owing to the changes of modulation sensitivity at the extremes of the characteristic, which changes are also significant with regard to the bandwidth of the feedback amplifier. The larger the modulation depth, the greater is the tendency for the transmission to remain constant after the reversal of rate of change occurs, resulting in rapid changes being demanded of the amplifier and involving much faster transients than those of the driving waveform. The current-drive waveform for 50% pure sine-wave modulation on the characteristics of Fig. 2B is shown in Fig. 3, together with the required drive waveforms for modula-

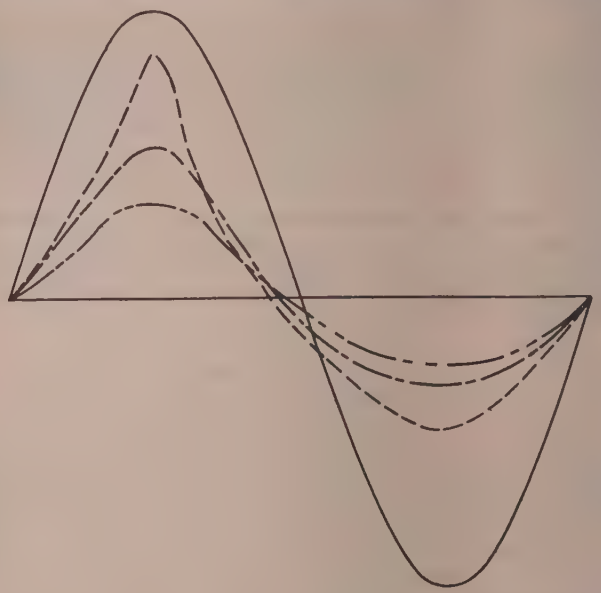


Fig. 3.—Current-drive waveforms required for pure sine-wave modulation at different modulation factors.

——— Reference sine-wave.
– – – – 100% modulation.
—·— 66% modulation.
—··— 50% modulation.

tions of 66% and 100% on similar characteristic cycles, which were determined experimentally.

Fig. 4 shows the basic feedback circuit from which a full circuit was developed. Briefly, the linear detector presents a waveform, representing the modulation envelope of the microwave signal, to a comparator, whose output is the instantaneous difference between this envelope and a reference waveform which represents the required modulation and is also fed to the comparator. The output is amplified and fed to the modulator magnetizing coil in the correct sense to cancel the difference at the comparator. There is a basic similarity between this process and envelope feedback as commonly employed for radio transmitters, with one important difference, namely that the final stage of the differential amplifier works into an inductive load which possesses hysteresis and is appreciably non-linear. Indirect feed to the comparator involves a detector whose linearity imposes the ultimate limitation on performance.

The choice of detector is important. Among the possibilities are a peak-rectifier type, a crystal operating on the linear part of its characteristic, and a superheterodyne detector involving a microwave local oscillator and an intermediate-frequency amplifier. It has been found that, of these, crystals possessing a range

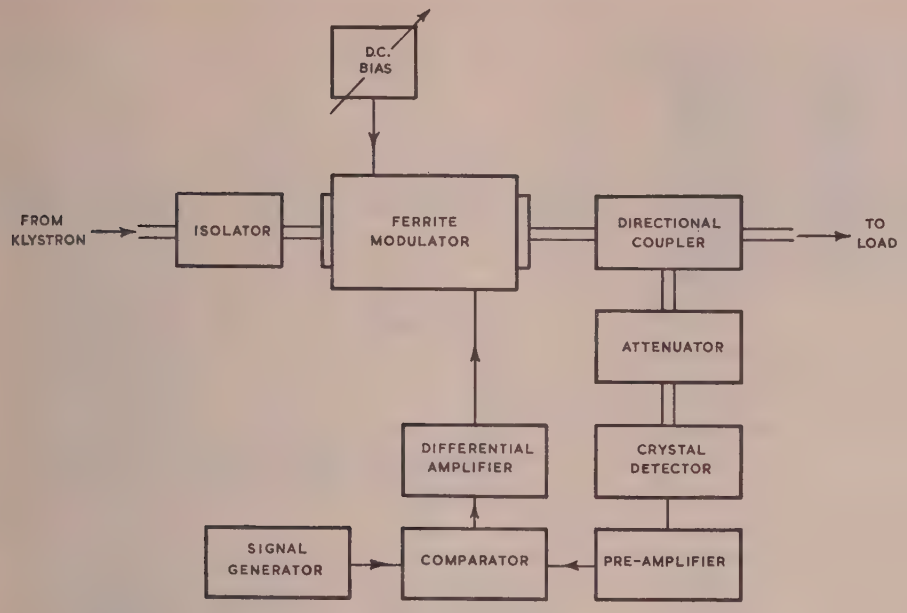


Fig. 4.—Block schematic of basic feedback circuit.

of linear characteristic may be used, and Fig. 5 shows such a characteristic measured with the aid of a precision attenuator. There is every indication that the non-linearity is substantially less than 0·5% over the marked portion of the curve, which permits 66% modulation depth. The attenuator in the feed to the crystal (Fig. 4) is used to set the detector operating-point, thus allowing the modulator operating-point to be optimized for the required depth of modulation. Crystal detectors are increasingly noisy with decreasing frequency (about 6 dB per octave is believed to be a representative figure); hence the crystal used in its linear range injects appreciable noise into the differential amplifier. Noise from the crystal behaves as if it were injected as reference signal to the comparator, and it is not cancelled by the servo action. Thus the loop gain and bandwidth which may be employed are noise-limited. In practice, it was found that the amplifier bandwidth needed to be restricted below the modulator response in order to employ a loop gain of 50 dB. This is partly because of phase distortion introduced when the magnetizing field penetrates the guide wall and partly because of the noise limitation. The feedback-amplifier phase characteristic might be arranged to compensate for the modulator over a band of frequencies and thus give less noisy linear detection. However, the modulator was empirically optimized for minimum power drive to provide a required transient response of less than one microsecond, and it seems that, for wideband feedback applications, improvement should properly concern the modulator and its driving stage. In the present application, feedback amplifier gain and bandwidth, within the limitations of the modulator response, the loop delay of the amplifiers and the noise factor of the detector, were satisfactory. A useful bandwidth of 100 kc/s was achieved with a loop gain of 50 dB, which indicates on the simplest feedback assumptions that unwanted voltages which appear have been suppressed to the extent of 1/300.

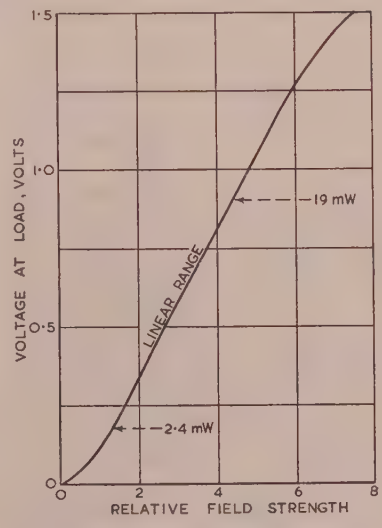


Fig. 5.—Crystal characteristic with load resistance of 300 ohms.

(4) FEEDBACK CIRCUITS AND PERFORMANCE

In practice it was found convenient to amplify the detected envelope before comparison with the reference (driving) waveform, and a test bench was set up, as shown in Fig. 6. Response curves of crystals and the static plots of modulation cycles were taken by this means.

These measurements led to the design of an amplifier for the error signal, shown in Fig. 7, while a standard wide-band pre-amplifier was employed between the crystal detector and the comparator. An analysis of the frequency response is indicated in Fig. 8, which gives the responses of the pre-amplifier, of the error-signal amplifier as far as the grid of the modulator driving valve, of the driving valve itself, and of the modulator magnetization. The driving valve response was obtained by applying constant voltage at the grid and measuring the modulator-coil current by the potential across a small resistor in series with it. The modulator response was obtained by applying constant small currents to the coil and measuring the amplitude of modulation detected by the crystal; the effects of hysteresis are fairly small

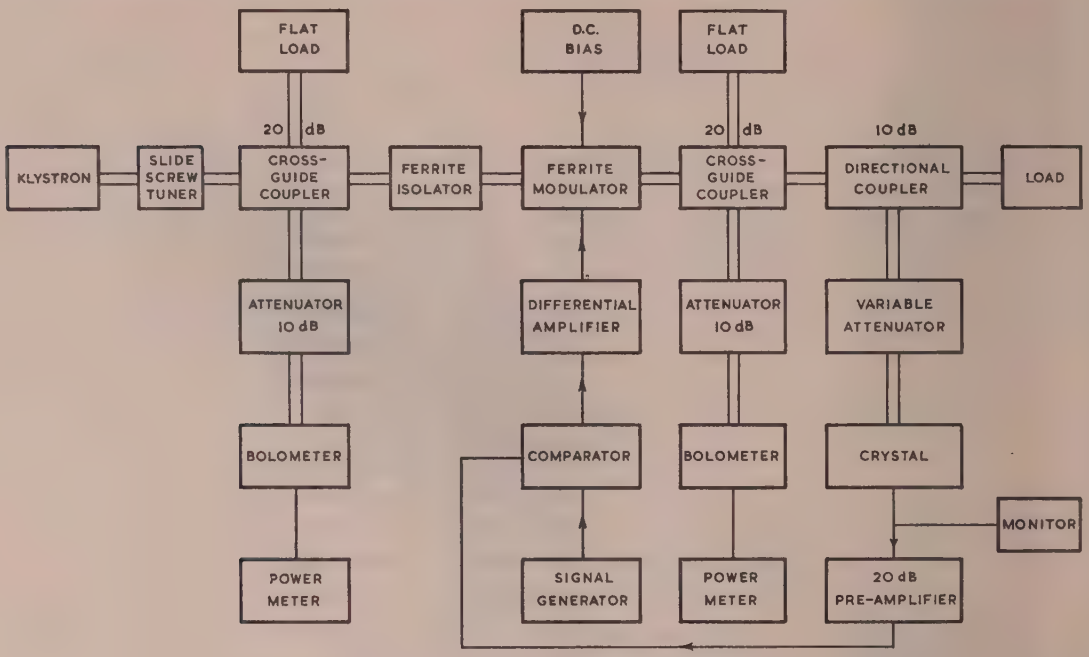


Fig. 6.—Block schematic of test bench arrangement.

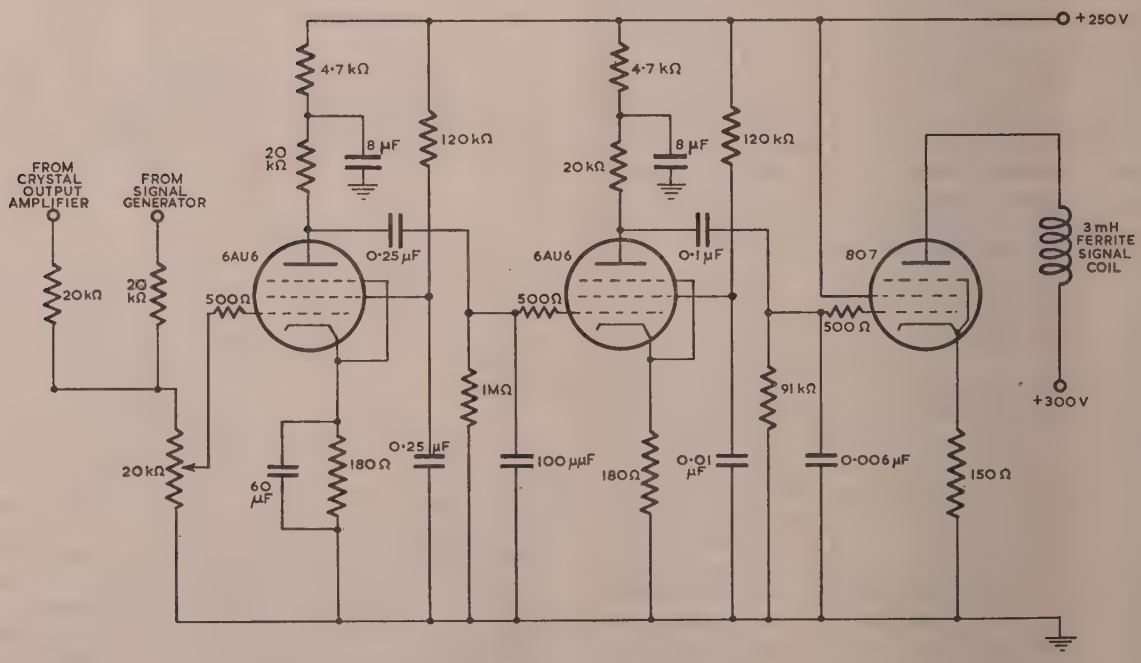


Fig. 7.—Comparator and differential amplifier.

for such a low-amplitude signal, and should not vary with frequency in the range concerned. Since the crystal works into a resistance of 330 ohms, its frequency response may be assumed not to affect the measurements. These two responses both depend on the design of the modulator. The loading of the coil and its tuned frequency are functions of the wall of the circular guide, the ferrite and the capacitance shunting the coil, while the magnetizing field produced in the ferrite is modified by eddy currents in the guide wall. Hence the modulator design provides the ultimate limit to frequency response (e.g. the 6 dB drop in the response between 200 c/s and 20 kc/s is believed to be caused by eddy currents in the flanges).

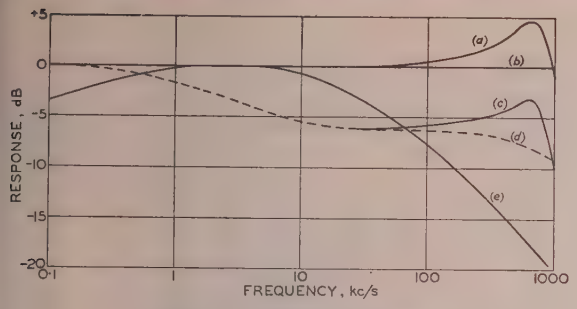


Fig. 8.—Frequency responses.

(*a*) Driving stage.
(*b*) Pre-amplifier.
(*c*) Modulator and driving stage.
(*d*) Modulator.
(*e*) Differential amplifier.

Within its frequency limitations, the modulation performance is satisfactory for many purposes, which may be seen from Figs. 9, 10 and 11. With the sinusoidal waveforms of Fig. 9, feedback produces a substantial reduction in harmonic content, as shown

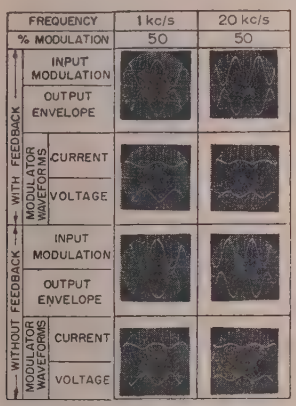


Fig. 9.—Actual waveforms. Sinusoidal functions, 50% modulation.

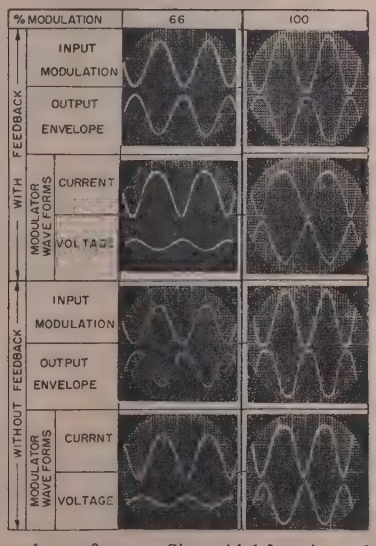


Fig. 10.—Actual waveforms. Sinusoidal functions, 66% and 100% modulation, at 1 kc/s.

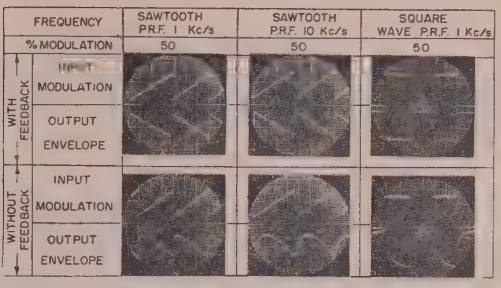


Fig. 11.—Actual waveforms. Non-sinusoidal functions.

in Table 1. The fidelity is consistent with the overall loop gain. Taking the modulation to higher percentages than are allowed by the crystal characteristic is also justified, as a test of feedback performance, since the feedback circuit still has to carry second

Table 1

HARMONIC ANALYSIS—SINUSOIDAL FUNCTIONS

Fundamental frequency	Waveform analysed	Modulation	Level below fundamental		
			2nd harm.	3rd harm.	4th harm.
kc/s		%	dB	dB	dB
1	Reference waveform (input modulation)		50	57	75
	Output envelope with feedback	25	49	53	75
		50	44	49	60
		66	41	50	73
		75	39	42	54
	Output envelope without feedback	25	27	38	64
		50	17	36	45
		66	26	33	57
		75	21	25	40
10	Reference waveform (input modulation)		55	53	80
	Output envelope with feedback	25	51	52	79
		50	46	47	67
	Output envelope without feedback	25	36	43	55
		50	34	35	50
20	Reference waveform (input modulation)		52	55	85
	Output envelope with feedback	25	52	38	58
		50	28	33	52
	Output envelope without feedback	25	39	35	52
		50	28	33	52

and higher harmonics in order to produce a detected sine wave. Fig. 10 shows the waveforms obtained at 66% and 100% modulation; the former represents the limit of crystal linearity. The harmonic content analysis of the reference waveform and the output envelopes with and without feedback for 66% and 75% modulation has been included in Table 1. The performance of the system may be better appreciated visually from consideration of Fig. 11, which relates to non-sinusoidal functions. Their harmonic analysis is presented in Table 2.

When feedback was not in use the amplifier gain was reduced to give the same percentage modulation. The photographed waveforms do not reveal the exact harmonic contents, but the current-drive waveforms with feedback, and comparison of the

Table 2

HARMONIC ANALYSIS—NON-SINUSOIDAL FUNCTIONS AT 50% MODULATION

Fundamental frequency and waveform	Waveform analysed	Level below fundamental										
		2nd	3rd	4th	5th	6th	7th	8th	9th	10th	19th	20th harm.
		dB	dB	dB	dB	dB	dB	dB	dB	dB	dB	dB
1 kc/s sawtooth	Reference waveform (input modulation)	6	10	12	14	16	17	18	20	21	—	27
	Output envelope with feedback ..	6	10	12	14	16	17	19	20	21	—	28
	Output envelope without feedback ..	7	11	14	17	19	21	23	25	26	—	36
10 kc/s sawtooth	Reference waveform (input modulation)	6	9	12	14	15	16	17	19	20	—	27
	Output envelope with feedback ..	6	9	12	14	16	18	20	22	23	—	38
	Output envelope without feedback ..	8	14	19	23	27	30	33	35	38	—	54
1 kc/s square-wave	Reference waveform (input modulation)	75	9	54	14	54	17	54	19	60	27	—
	Output envelope with feedback ..	39	10	42	14	44	18	45	20	47	27	—
	Output envelope without feedback ..	21	12	28	18	32	22	36	25	38	35	—

detected envelopes with and without feedback, clearly reveal the distortions which are corrected by the feedback amplifier.

It was stated earlier that square-wave modulation may be applied without appreciable distortion; this is in fact qualified by the requirement to hold the coil current levels, at top and bottom, very closely. Otherwise, distortion occurs without feedback, small imperfections of the square-wave being strongly emphasized, as seen in Fig. 11.

Finally, it is necessary to consider the importance of linearity in the amplifier and modulator driving stages. Since the modulator itself is definitely non-linear, it is required only that the linearity of the overall loop be not greatly changed by that of the amplifier. It is well known that the exact statement of behaviour in non-linear devices is difficult, but feedback is a standard method of producing overall linearity, and has been frequently described analytically for roles similar to that here discussed.

It is of interest to note that the addition of the feedback circuit had the effect of suppressing all amplitude-modulation noise generated by the klystron in the pass band of the feedback circuit.

(5) CONCLUSION

The microwave modulator and associated feedback circuit which have been described represent a first approach towards the employment of ferrite devices for linear modulation. In order to achieve further linearity, a necessary step is the provision of a detector with a wider linear range, and to justify a higher loop gain a less noisy detection circuit would be necessary. Then the provision of further feedback gain with special precautions to reduce the loop delay would be justified. There is no *a priori* reason why the design principles should not be repeated for higher frequencies, where similar results might be achieved.

Employing envelope feedback, sine waves have been impressed on a microwave signal at frequencies up to 20 kc/s and faithfully reproduced with harmonic sidebands more than 45 dB below the fundamental. Non-sinusoidal functions give better visual indication of the feedback performance, and the comparative results presented for both square and sawtooth waves underline the value of the feedback method.

(6) ACKNOWLEDGMENT

The authors are indebted to the Defence Research Board of Canada for permission to publish the paper. They wish to express their appreciation of helpful discussions with Mr. E. A. Walker and Mr. P. M. Thompson.

(7) REFERENCES

(1) MILLER, T.: 'Magnetically Controlled Waveguide Attenuators', *Journal of Applied Physics*, 1949, **20**, p. 878.

(2) HOGAN, C. L.: 'The Microwave Gyrator', *Bell System Technical Journal*, 1952, **31**, p. 1.

(3) SAKIOTIS, N. G., SIMMONS, A. J., and CHAIT, H. N.: 'Microwave-antenna Ferrite Applications', *Electronics*, June, 1952, **25**, p. 156.

(4) REGGIA, F., and BEATTY, R. W.: 'Characteristics of the Magnetic Attenuator at UHF', *Proceedings of the Institute of Radio Engineers*, January, 1953, **41**, p. 93.

(5) OLIN, I. D.: 'An X-Band Sweep Oscillator', *Proceedings of the Institute of Radio Engineers*, January, 1953, **41**, p. 10.

(6) BARRY, J. N., and CLARKE, W. W. H.: 'Microwave Modulator uses Ferrite Gyrator', *Electronics*, May, 1955, **28**, p. 139.

(7) CACHERIS, J.: 'Microwave Single-Sideband Modulator Using Ferrites', *Proceedings of the Institute of Radio Engineers*, August, 1954, **42**, p. 1242.

(8) KALES, M. L., CHAIT, H. N., and SAKIOTIS, N. G.: 'A Non-Reciprocal Microwave Component', *Journal of Applied Physics*, 1953, **24**, p. 816.

(9) ALBERS-SCHOENBERG, E.: 'Ferrites for Microwave Circuits and Digital Computers', *Journal of Applied Physics*, 1954, **25**, p. 152.

621.387:621.372.43

Paper No. 2060 R
July 1956

WIDE-BAND NOISE SOURCES USING CYLINDRICAL GAS-DISCHARGE TUBES IN TWO-CONDUCTOR LINES

By R. I. SKINNER, B.E.

(*The paper was first received* 1*st January*, 1955, *and in revised form* 31*st January*, 1956.)

SUMMARY

The provision of wide-band noise sources for the decimetre wavelength region is important for many radio applications. A noise signal of suitable power level for most purposes is obtained when the plasma region of a gaseous discharge is matched to a transmission line.

Examination of the properties of a discharge plasma shows that a noise source with an output which is constant over several octaves can be obtained by matching a cylindrical discharge tube directly to a two-conductor line. Such matching can be achieved by using conductor pairs of various shapes. The factors which affect the operation of the matching element are considered, and a practical design procedure is outlined.

These noise sources are simpler to construct and of better performance than those used previously at decimetre wavelengths.

LIST OF SYMBOLS

Z_0 = Free-space line impedance.
ρ_s = Area resistivity.
R = Resistance.
D, d = Diameter.
Z = Complex impedance of the line.
L = Line inductance per unit length.
C = Line capacitance per unit length.
γ = Complex propagation coefficient of the line.
ω = Angular frequency.
μ_0 = Free-space permeability.
χ = Electric susceptibility.
ϵ = Relative permittivity.
ϵ_0 = Absolute permittivity.
σ = Complex conductivity.
X_c = Capacitive reactance.
ϵ_a = Average absolute permittivity.
l = Electron mean-free-path.
λ = Wavelength.
T = Electron absolute temperature.
n = Electron density.
A = Internal radius of discharge tube.
r = Distance from axis of discharge tube.
n_r = Electron density at distance r.
J = Average discharge current density.
v = Electron axial drift velocity.
C' = Complex line capacitance per unit length.
X_c' = Complex capacitive reactance.
$X_c'(\Delta s)$ = Complex capacitive reactance across elementary prism.

The rationalized system of MKS units is employed throughout the paper.

(1) INTRODUCTION

Investigations in such fields as radio astronomy and the development of more sensitive radio receivers require the accurate measurement of small noise signals. Such noise signals may be compared directly with a reference signal of white noise which is therefore more convenient than a reference signal with a discrete frequency.

A white-noise signal with a power level suitable for many sensitivity measurements is provided by a temperature-limited diode, which may be used conveniently at wavelengths down to one metre. At shorter wavelengths, however, transit-time effects reduce the noise signal output by an amount which is difficult to assess.

It was shown by Mumford[1] that the plasma region of a gaseous-discharge tube generates a white-noise signal of 15–21 dB above ambient thermal-noise level, a power adequate for radio measurements in the microwave region. The noise signal was demonstrated by placing the plasma region of a discharge tube across a waveguide and matching the plasma to the waveguide by means of reactance plugs. A more practical form of the gas-discharge noise source was developed by Johnson and Deremer,[2] who placed a discharge tube diagonally across a waveguide to obtain a relatively wide-band noise source, which was matched to the waveguide over its operating bandwidth.

As the wavelength is increased above 10 cm waveguide techniques give way to more compact transmission systems based upon two-conductor lines. A two-conductor-line noise source is desirable for use at these longer wavelengths, and such an instrument was developed by Johnson,[3] who made use of the inherently large bandwidth of a two-conductor line to produce a wide-band noise source covering about one and a half octaves. The gradual introduction of the discharge necessary for a wide-band match was achieved by means of a tapered discharge tube. While a bandwidth of 10–30 cm was obtained with a matching element 18 in long, two disadvantages remained. The expendable discharge tube was difficult to construct, and a more serious disadvantage was the variation of noise power which must occur throughout the operating wavelength range. This arises because the level of the noise signal varies with the tube diameter and this varying signal is sampled differently at different wavelengths.

Since a discharge plasma is a lossy material it was natural that in the early development of gaseous-discharge noise sources the system of matching should follow by direct analogy with previously used transmission-line terminations employing lossy materials. In Fig. 1 a comparison is made between the earlier noise sources and the corresponding line terminations. However, too strict an adherence to these techniques leads, with the wide-band two-conductor noise source of Fig. 1(*f*), to an undesirable variation in noise power with the wavelength. Since a cylindrical discharge tube, when matched to a passive load, generates a noise power independent of the wavelength,[2] a more satisfactory noise source would result if such a discharge were matched directly to a two-conductor line. A cylindrical discharge can in fact be matched over a wavelength range of several octaves. To obtain a match, the two conductors are distorted, as shown in Fig. 2, to introduce the cylindrical discharge gradually, and hence give a wide-band match.

The form shown in Fig. 2(*a*), where the discharge tube is matched to a coaxial line, is particularly suitable for application at decimetre wavelengths, but the forms shown in Figs. 2(*b*) and

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.
Mr. Skinner is in the Dominion Physical Laboratory, D.S.I.R., New Zealand.

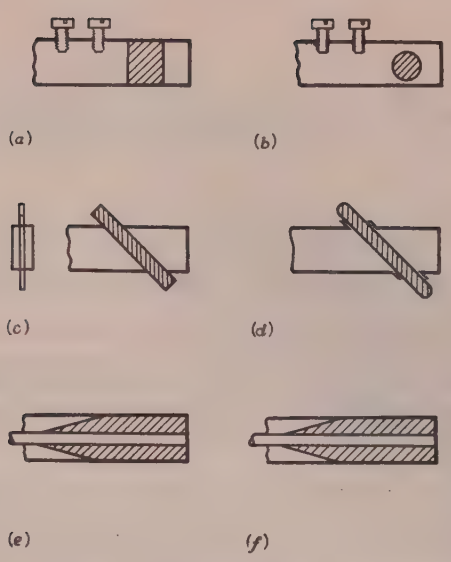


Fig. 1.—Matched terminations with lossy materials and with discharge tubes.

(*a*) Waveguide with narrow-band lossy-material termination.
(*b*) Waveguide with narrow-band discharge-tube termination (Mumford).
(*c*) Waveguide with wide-band lossy-material termination.
(*d*) Waveguide with wide-band discharge-tube termination (Johnson and de Remer).
(*e*) Coaxial line with wide-band lossy-material termination.
(*f*) Coaxial line with wide-band discharge-tube termination (Johnson).

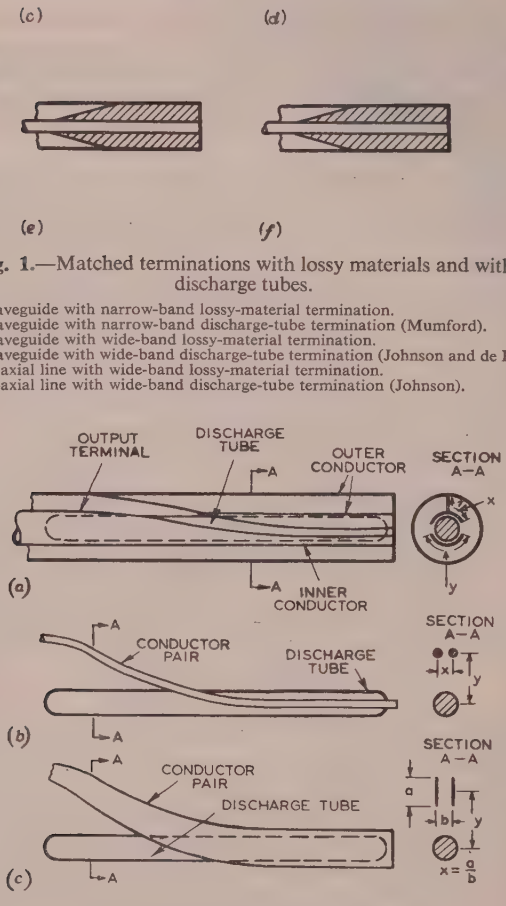


Fig. 2.—Two-conductor matching elements.

(*a*) Closed line with low impedance.
(*b*) Open line with high impedance.
(*c*) Open line with low impedance.

2(*c*) should prove suitable for some applications at longer wavelengths. The following discussion of design considerations applies specifically to the coaxial-line noise source of Fig. 2(*a*), but with some minor modification it could be applied to the open-line type.

In the coaxial-line noise source the discharge tube is placed within the hollow inner conductor, from which is cut a segment of increasing size to expose an increasing amount of discharge plasma to the line. The increase in the line impedance which results from the removal of part of the inner conductor is offset by an appropriate addition to the outer conductor. The blade and fin added to the outer conductor also increase the electric field intensity across the discharge plasma. In order to control the line propagation characteristics along the matching element to give a wide-band match, some means of assessing these characteristics must be found. Because the conductor geometry is complex, and since the dielectric properties vary over a line cross-section, the accurate determination of the transmission-line characteristics of the matching element is an extremely laborious process. However, an approximate method making use of experimental data has been developed, and has proved highly satisfactory for designing wide-band noise sources for the decimetre region.

In Section 2, methods for the determination of the line propagation characteristics of the matching element are discussed. The use of these propagation characteristics in designing a matching element are considered in Section 3. In Sections 4 and 5 the approximate design and operating characteristics of two experimental noise sources are described.

(2) DETERMINATION OF THE LINE PROPAGATION CHARACTERISTICS

In the design of any matching element, it is necessary to consider three interrelated parameters: the wavelength range, the maximum voltage standing-wave ratio, and the overall length. When any two of these parameters are given, the optimum value for the third is defined, the limitation being the values of line propagation characteristic which can be obtained. However, no design procedure has been developed which gives this best possible matching element. It is shown in Section 3 that a practicable design procedure, which results in an effective but not optimum matching element, can be based upon a knowledge of the line impedance along the matching element for free-space dielectric conditions and of the line impedance at one cross-section containing the discharge tube.

(2.1) Free-Space Line Impedance

The free-space line impedance may be obtained by plotting the flux of the TEM fields[8] or from measurements on an electrolytic tank analogue. The electrolytic tank must provide a sheet of uniform resistivity between a pair of conductors which are shaped to correspond to the cross-section of the transmission line at the point at which the impedance is to be obtained. This may be achieved conveniently by using cylindrical model conductors, the cross-section of which has the same shape as that of the transmission line. The conductors are placed vertically in an electrolyte of uniform depth. If a low-frequency voltage is applied between the model conductors, the potential and current-flux net has the same shape as the flux net of the transverse electromagnetic waves in the transmission line, since both are derived from potentials which satisfy Laplace's equation and have the same boundary conditions, if the effect of resistivity in the line conductors is neglected. The resistivity of the electrolyte sheet in the model replaces the wave impedance of free space in the line, and hence the measured resistance between the model conductors is proportional to the impedance which the transmission line presents to transverse electromagnetic waves, so that

$$Z_0 = 377\frac{R}{\rho_s} \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

where Z_0 = free-space line impedance.
ρ_s = resistivity of the sheet of electrolyte.
R = resistance between the model conductors.

The value of ρ_s may be obtained easily if a pair of conductors, of diameters D and d, are immersed concentrically in the electrolyte, since the resistance between them is

$$R = \frac{\rho_s}{2\pi}\log_e(D/d)$$

(2.2) Line Propagation Characteristics with the Discharge Tube Present

In the following discussion it is convenient to take into account the conductance G across a capacitor by thinking in terms of a complex capacitance C' and a corresponding complex reactance C' is given in terms of the capacitance C, as usually defined,

$$C' = C + G/j\omega$$

For any point along the matching element of Fig. 2(*a*) the line impedance and propagation coefficient may be expressed as

$$Z = \sqrt{(L/C')} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

$$\gamma = j\omega\sqrt{(LC')} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (3)$$

When evaluating the line inductance and capacitance for substitution in eqns. (2) and (3) the following simplifying assumptions are made. The line inductance is taken as that which is presented to TEM fields propagated under free-space conditions between the conductors. The line capacitance is taken as the static capacitance which exists between the conductors when the distribution of the permittivity is the same as that encountered by electromagnetic fields of angular frequency ω. From simple physical considerations it is evident that neither the electric field nor the magnetic field is of the form assumed, but the error involved in making these assumptions is difficult to estimate. However, an experimental check was made by measuring, at wavelengths from 20 to 30 cm, the propagation coefficient of a line partially filled with dielectric. This measured propagation coefficient agreed with the value given by eqn. (3), when using the above approximations for the line inductance and capacitance, to within the limits of error of the experiment, i.e. $\pm 2\%$. Line impedance values based upon the approximate capacitance and inductance should have comparable accuracies.

While the propagation characteristics are obtained for a given point on the tapered matching element, it is convenient, when discussing the corresponding inductance and capacitance, to think in terms of a uniform line of unit length with the same cross-section as the matching element at the point considered. This concept is used throughout the remainder of Section 2 and in Section 3.

(2.2.1) Free-Space Line Inductance.

The free-space transmission-line inductance can be obtained from the electrolytic tank analogue described in Section 2.1 in connection with free-space line impedance. The resistance between the model conductors immersed in a uniform depth of electrolyte is measured as before. The equipotentials of voltage in the electrolyte now correspond to lines of magnetic flux in the transmission line, so the resistance measured is inversely proportional to the reluctance in the magnetic flux path and hence directly proportional to the line inductance. The free-space inductance per unit length of line is therefore given by

$$L = \frac{\mu_0 R}{\rho_s} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (4)$$

The value of ρ_s may be obtained as before.

(2.2.2) Static Capacitance of the Line.

Two steps are required to obtain the static capacitance between the conductors per unit length of transmission line. The permittivity is obtained at all points over the cross-section of the line, and in particular throughout the area occupied by the discharge plasma. Unit length of line is then represented by an equivalent circuit, as described by Kron.[7] From the overall impedance of this network the line capacitance is obtained.

The permittivity of the discharge plasma depends upon the complex conductivity of its electron cloud. The permittivity is best expressed in terms of the electric susceptibility, since the susceptibility always varies in the same way over a cross-section of the discharge tube. The relation is

$$\chi = \epsilon - 1 = \frac{j\sigma}{\epsilon_0} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (5)$$

A formula has been developed by Margenau[4] which gives the conductivity of the plasma in terms of the electron density, temperature and mean free path, and the wavelength. These parameters may be obtained from the literature, and the conductivity evaluated as described in Section 8.

Fig. 3 gives values for the electric susceptibility on the axis of a discharge in neon carrying a direct current of 25 mA in a tube

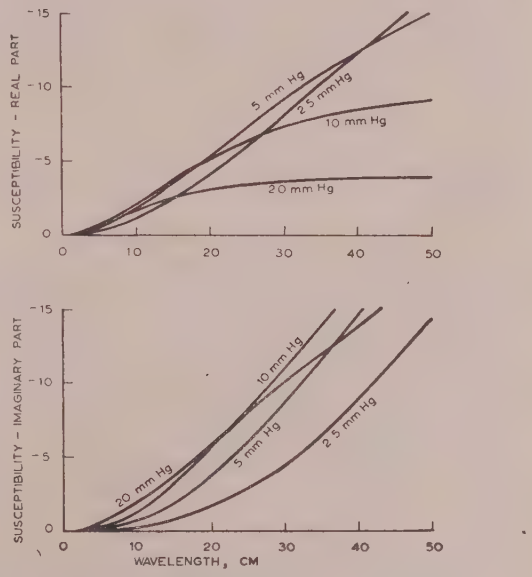


Fig. 3.—Electric susceptibility on the axis of a discharge tube plasma.

The curves are for a discharge tube of 1·2 cm internal diameter filled with neon.

of 1·2 cm diameter. The real and imaginary parts of χ are proportional to σ by eqn. (5). It is shown in Section 8 that σ varies according to a zero-order Bessel function with the origin on the tube axis and the first zero at the tube walls. The relative permittivity at any point across the discharge plasma follows from σ by the second part of eqn. (5). The relative permittivity is known for the glass wall of the discharge tube and for the air space, so it is now known throughout the cross-section of the transmission line. The line capacitance may now be calculated as described below.

The reactance of the capacitance between unit length of the two transmission-line conductors may be represented, for a given wavelength, by a network of impedances as described by Kron.[7] To set up this equivalent network the dielectric between the conductors is first divided into elementary rectangular prisms by three sets of orthogonal surfaces. It is convenient to consider filamentary prisms terminated by the end planes. The remaining two sets of surfaces are parallel to the line axis and intersect the end planes in the rectilinear squares of a free-space electric-flux and equipotential plot. The dielectric is thus divided into elementary prisms of unit length and square cross-section. The

impedance which a prism offers to the flow of electric flux between a pair of opposite long faces is

$$X'_c(\Delta s) = -\frac{1}{\epsilon_a} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (6)$$

The overall capacitive reactance between the conductors can now be obtained by connecting these reactance elements into a rectangular array and solving the matrix equation of the network. The capacitance between unit length of the conductors is then

$$C' = -\frac{1}{\omega X'_c} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (7)$$

This value of the line capacitance together with the line inductance obtained by electrolytic tank measurements may be substituted in eqns. (2) and (3) to obtain the line propagation characteristics.

The detailed determination of the line propagation characteristics described above is unwieldly in practice since it involves the solution of a high-order matrix equation, but might be practicable where suitable computer facilities are available. However, the method of solution obtained above illustrates clearly the interrelation of all the factors involved. It also points towards and provides a check for approximate methods of solution.

(2.3) Approximate Line Capacitance

The line capacitance may be obtained approximately at the longer wavelengths to be matched by assuming that the part of the cross-section occupied by the discharge tube and the inner and outer conductors in contact with it is a partial cylindrical condenser. At these longer wavelengths the central region of the discharge plasma has a high permittivity, as shown by the curves of Fig. 3, and this central region may therefore be replaced by a cylindrical conductor with little loss of accuracy. The capacitance of this condenser may be obtained by a straightforward integration along a radial line, together with an appropriate allowance for fringing fields. In a trial calculation for a typical instance it was found that the amplitude and phase of the capacitance obtained from the radial condenser approximation differed by less than 10% from the capacitance obtained over the same region by the more detailed equivalent circuit of the Kron method. The capacitance of the air-space area between the conductors is added to the radial condenser capacitance to obtain the total line capacitance per unit length for the given cross-section. The line propagation characteristics are then obtained as before from eqns. (2) and (3).

(3) SATISFYING THE CONDITIONS FOR A WIDE-BAND MATCH

The conditions which must be satisfied to achieve a wide-band match to a lossy material follow from straightforward transmission-line theory. They are:

(*a*) The change of the line impedance must be small over a distance comparable with the wavelength.

(*b*) Adequate overall attenuation must be introduced.

The first of these conditions is satisfied by keeping the amplitude of the line impedance as nearly constant as possible along the matching element, at the longest wavelength to be matched. It should be noted that changes in the cross-section of the conductors can control the amplitude but not the phase of the line impedance. The inner conductor is opened out to introduce the discharge tube gradually over a length of matching element equal to about twice the longest wavelength. This gradual introduction limits reflections due to the unavoidable changes in the phase of the line impedance. As the wavelength is reduced the plasma permittivity falls and the changes in the phase of the line impedance are generally smaller. At the same time, the line impedance increases and hence counters the effect of the reduced phase changes, which taken alone tend to improve the match. However, any increased changes in the value of the line impedance are largely offset by the increase in the electrical length of the matching element at the shorter wavelength.

As the wavelength is decreased the attenuation introduced by the discharge plasma falls until the second condition for a wide-band match is no longer satisfied and excessive reflections are returned from the end of the line. The shortest wavelength matched may be reduced by increasing the overall length of the matching element or redesigning the taper, using a plasma which gives a higher attenuation. The above may be compared with the common method of forming a wide-band line termination by placing a wedge of lossy material in a uniform continuation of a line. This corresponds to noise sources in which the free-space line impedance along the matching element is held constant, a condition which may be achieved by the use of results from the electrolytic tank analogue. With some line terminations the impedance may be changed gradually along a wedge of lossy material in an attempt to offset, as far as possible, the impedance changes due to the lossy material, and these correspond to the noise sources described below, in which the conductor geometry which gives a constant free-space impedance is modified to offset the impedance changes due to the discharge tubes.

(4) EXPERIMENTAL NOISE SOURCES

A noise source of the form shown in Fig. 2(*a*) was constructed for the 10–35 cm wavelength range. The steps in its development follow from the general considerations of the previous Section.

The coaxial cylinders upon which the conductor pair was based correspond to a 53·5-ohm line. A model of this conductor pair was constructed which could be adjusted readily to give any of the possible forms of inner and outer conductor. The free-space line impedances were determined by substituting in eqn. (1) resistance values obtained from measurements on the model conductors, which were immersed in an electrolyte. These impedances are given by the curves in Fig. 4.

A discharge tube of internal diameter 1·2 cm, filled with neon to a pressure of 10 mm Hg and operated with a direct current of 30 mA, was chosen as likely to provide sufficient attenuation at a wavelength of 10 cm. The permittivity across this discharge was calculated at 35 cm by applying eqn. (5) to the complex conductivity as determined by the method outlined in Section 8.

The inner conductor of the transmission line was cut away to expose an increasing amount of discharge tube until only one-eighth of its circumference remained. The shape of the outer conductor necessary to maintain the free-space line impedance at 53·5 ohms throughout the taper was obtained from Fig. 4. The line cross-section giving maximum exposure of the discharge tube was then examined to assess the effect of the tube on the line impedance. The line capacitance was obtained by using the radial condenser approximation described in Section 2.3, and the line inductance by applying eqn. (4) to the electrolytic tank measurements. The line impedance was then obtained by substitution in eqn. (3), and it was found that the discharge tube reduced the line impedance to about two-thirds of its free-space value, for the condition of maximum tube exposure. The data in Fig. 4 were used to determine the shape which must be given to the outer conductor to offset this impedance reduction. In the corrected taper the line impedance increased gradually from 53·5 ohms to 80 ohms, as the inner conductor was cut away.

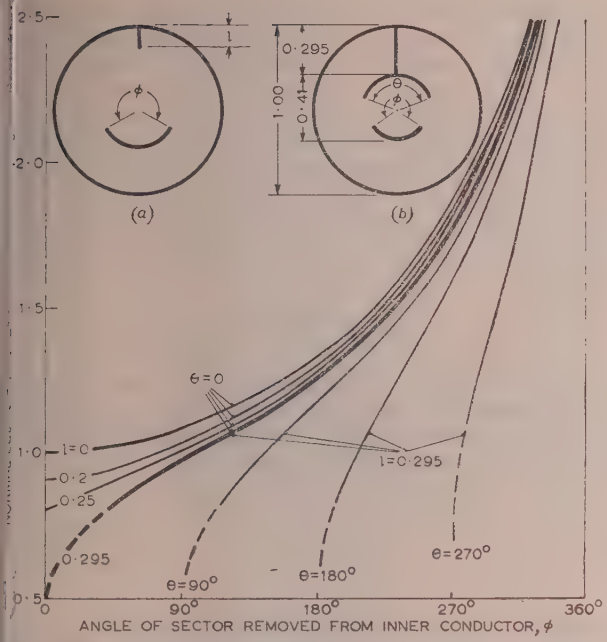


Fig. 4.—Free-space characteristic impedance for various cross-sections of the two-conductor line upon which the matching element is based.

l = Width of blade added to outer conductor.
θ = Angle of sector added to blade on outer conductor.
φ = Angle of sector removed from inner conductor.
(*a*) Cross-section of line in which $\varphi = 0$.
(*b*) Cross-section of line in which $l = 0{\cdot}295$.

An experimental noise source was then constructed in which the tapered conductors extended over a length of 14 in. Beyond the taper the conductors were continued uniformly with the maximum exposure of the discharge tube, for a further 14 in. An adjustable short-circuit was placed in this part of the line, and by this means it was established that a matching element with a minimum total length of 18 in was required to provide adequate overall attenuation for match down to a wavelength of 10 cm.

A second experimental noise source was built to show that a more compact structure with a correspondingly reduced wavelength range could be obtained. In this noise source the reflection from the steeper tapered part of the conductors was balanced against reflections from the end of the line, the limited wavelength range being a characteristic of such a balance. The particular wavelength range obtained can be varied considerably by moving the position of the terminating short-circuit and at the same time changing the discharge current.

An outline of the operating conditions and performance of the two experimental noise sources is given in Table 1.

Table 1

RANGE AND PERFORMANCE OF EXPERIMENTAL NOISE SOURCES

	Model I	Model II
Noise power (approx.)	17 dB	17 dB
Wavelength range	9–37 cm	22·5–24·5 cm
Maximum v.s.w.r.	1·15	1·05
Length of matching element ..	24 in	7 in
Gas used	Neon	Neon
Discharge-tube pressure	12 mm Hg	12 mm Hg
Discharge-tube internal diameter	12 mm	12 mm
Discharge-tube current	30 mA	30 mA

It was not possible to measure the noise power precisely with available facilities. However, the output is independent of the wavelength if two conditions are fulfilled:

(*a*) The available noise power from a cylindrical discharge tube must be independent of the wavelength. This independence has been verified with considerable accuracy.[2]

(*b*) The discharge plasma must be matched to its load. The effectiveness of the matches obtained is indicated by the values of the v.s.w.r. given in Table 1.

Some plasma instability is to be expected with the discharge-tube pressure and current density used in the experimental noise sources. While voltage fluctuations existed between the tube electrodes, no modulation of the output noise was detected.

More detailed investigations may prove that modulation exists on the noise output to an extent which is undesirable for some applications, while more accurate measurements may reveal longer-term variations. Such difficulties are common to all gas-discharge noise sources and may be removed if necessary by one of the three commonly used methods.[2] However, such modifications do not invalidate the design procedure described.

(5) CONCLUSIONS

The noise sources designed give a uniform noise output over a frequency range of four to one in the decimetre region. This range could be increased by a more extensive use of the design methods discussed. Careful design and manufacture should result in matching elements which are comparable with the best wide-band terminations obtained with other lossy materials.

The noise sources are comparatively simple to manufacture, while the expendable discharge tube is of the simplest form possible; it consists of a cylindrical discharge tube fitted with standard commercial electrodes and filled with the appropriate rare gas at a pressure of a few millimetres of mercury.

The straightforward design procedure outlined gives the conductor profiles for a matching element of any specified wavelength range.

(6) ACKNOWLEDGMENTS

Acknowledgment is made to Mr. W. H. Ward, Mr. G. W. G. Court and Mr. R. S. Unwin for useful discussions and other assistance during the preparation of the paper.

(7) REFERENCES

(1) MUMFORD, W. W.: 'A Broad-Band Microwave Noise Source', *Bell System Technical Journal*, 1949, **28**, p. 108.
(2) JOHNSON, H., and DEREMER, K. R.: 'Gaseous Discharge Super-High-Frequency Noise Sources', *Proceedings of the Institute of Radio Engineers*, 1951, **39**, p. 908.
(3) JOHNSON, H.: 'Super-High-Frequency Noise Source', R.C.A. Quarterly Report No. 15, 1951.
(4) MARGENAU, H.: 'Conduction and Dispersion of Ionized Gases at High Frequencies', *Physical Review*, 1946, **69**, p. 508.
(5) COBINE, J. D.: 'Gaseous Conductors' (McGraw-Hill, New York, 1941), pp. 23 and 235.
(6) LOEB, L. B.: 'Fundamental Processes of Electrical Discharges in Gases' (John Wiley and Sons, New York, 1939), p. 192.
(7) KRON, G.: 'Electric Circuit Models of Partial Differential Equations', *Electrical Engineering*, 1948, **67**, p. 672.
(8) BOOKER, H. C.: 'The Elements of Wave Propagation using the Impedance Concept', *Journal I.E.E.*, 1947, **94**, Part III, p. 171.

(8) APPENDIX

Discharge Plasma Conductivity

Margenau has developed an expression for the conductivity of a discharge plasma, valid for the conditions under which microwave noise sources are operated. In M.K.S. units the case of a plasma electron cloud reduces to

$$\sigma = 3{\cdot}82 \times 10^{-12} l n T^{-\frac{1}{2}} \left[K_2(x) - jx^{\frac{1}{2}} K_{3/2}(x)\right] \text{mho/metre} \qquad (7)$$

where

$$K_n(x) = \int_0^\infty \frac{y^n \varepsilon^{-y}}{x + y}\, dy$$

and

$$x = 1{\cdot}17 \times 10^{11} l^2 \lambda^{-2} T^{-1}$$

Values for $K_2(x)$ and $K_{3/2}(x)$ may be obtained most easily from the curves of Fig. 1 in Margenau's paper.[4]

For substitution in the above equation the electron temperature and mean free path may be obtained from standard texts.[5]

The electron density is obtained from the mean current density the charge distribution across the discharge tube, and the axia drift velocity of the electrons. It is shown by Cobine[5] that the charge distribution across a discharge tube (under gas pressure and current densities encountered in noise sources) varies accord ing to a zero-order Bessel function with the origin on the tube axis and the first zero at the tube wall. Since the electron mobility and the axial electric intensity are uniform over the cross section of a discharge plasma, the electron drift velocity is also uniform over the cross-section. It therefore follows that the charge density at any distance r from the axis of a discharge plasma is

$$n_r = 1{\cdot}45 \times 10^{19} \frac{I}{v} \mathrm{J}_0 (2{\cdot}40 r/A) \text{ electrons/cubic metre} \qquad (8$$

The axial drift velocity of the electrons in a discharge plasma may be obtained for helium, neon, and argon by comparing the curves published by Loeb and Cobine on pages 192 and 193 o Reference 6 and page 235 of Reference 5, respectively.

621.387.4 : 621.374.3 : 621.311.6 : 621.314.7

Paper No. 2049 M
Mar. 1956

THE APPLICATION OF TRANSISTORS TO THE TRIGGER, RATEMETER AND POWER-SUPPLY CIRCUITS OF RADIATION MONITORS

By E. FRANKLIN, Ph.D., Associate Member, and J. B. JAMES.

The paper was first received 7th October, 1955, and in revised form 27th January, 1956. It was published in March, 1956, and was read before a Joint Meeting of the MEASUREMENT AND CONTROL SECTION *and the* RADIO AND TELECOMMUNICATION SECTION *27th March, 1956.)*

SUMMARY

The paper outlines the general requirements and the conditions of use of radiation monitors employed in γ- and β-ray survey in connection with geological prospecting. In such instruments, it has been usual to employ either filament valves or cold-cathode valves in amplifier and trigger circuits, and vibrators, filament-valve oscillators or high-voltage battery stacks in the power supplies. Arguments leading to the transistor as the preferable component in all cases are given, and typical transistor circuits are discussed in some detail.

Both point-contact and junction transistors are discussed, and the superiority of the junction-type circuits for this type of application is demonstrated.

LIST OF PRINCIPAL SYMBOLS

I_{co} = Back current of transistor, collector to base with zero emitter current.
I'_{co} = Back current of transistor, collector to emitter with zero base current.
I''_{co} = Back current of transistor, collector to base and emitter.
v_{cb} = Voltage collector to base.
v_{ce} = Voltage collector to emitter.
v_t = Voltage across collector winding of transformer.
i_c = Collector current.
i_b = Base current.
i_e = Emitter current.
r_e = Emitter resistance.
r_b = Base resistance.
α' = Current gain of transistor (base to collector).

(1) INTRODUCTION

The circuits described in the paper were designed primarily for use in portable instruments intended for geological applications, such as uranium ore prospecting. Such instruments are subject to fairly rough usage, while being exposed to outdoor conditions in a wide range of climate. It is essential that they be reliable under such conditions, since they are frequently used in areas which are widely remote from servicing facilities or sources of spare parts. It is desirable also that any maintenance requirements, such as batteries, should be of types which are easily available anywhere in the world.

Such an instrument usually consists of a radiation detector, which is generally a Geiger–Müller counter or scintillation counter, a trigger circuit and a count-rate meter. The trigger circuit is triggered by the small pulses from the detector and produces larger pulses of uniform shape and amplitude independently of the form or size of the detector pulses. The count-rate meter is often merely a current meter with its associated integration circuit, which measures the value of a current produced by the trigger circuit proportional to the detector count rate or some simple function of it. The current produced by the trigger circuit is in the form of unidirectional pulses of a frequency which varies randomly about a mean value, and the purpose of the integrating circuit is to smooth this current to a sufficient degree to give a fairly steady indication on the meter. In the interests of power economy, it is desirable to keep the current as low as possible, but considerations of robustness and reliability set a limit to the sensitivity of the meter, and it is usual to employ a meter whose full-scale current is not less than 50 μA.

In some cases, particularly when the detector is a scintillation counter, there may be a pulse amplifier between the detector and the trigger circuit.

In the design of the trigger and rate-meter circuits, we are restricted by practical considerations in the choice of amplifying elements to the hot-cathode valve, the cold-cathode valve and the transistor. The properties of hot-cathode valves make circuit design a much easier problem than in the case of the two others. However, they are very uneconomical where battery power is concerned, since they require power for heating their cathodes, and also, in general, at least one valve in the circuit operates with an appreciable standing anode current. Also, in the experience of the authors they have proved less reliable than cold-cathode valves under the arduous conditions of use.

Cold-cathode valves have the great attraction of requiring no filament current and no standing h.t. supply apart from a very low 'keep alive' current of less than 1 μA, and their properties lead to extremely simple circuits in this type of application. They require a greater h.t. voltage (130–200 volts) than certain types of hot-cathode valve, but this is not a great disadvantage, because, in any event, means have to be provided for generating a high-voltage supply for the Geiger–Müller counter (350–500 volts), or scintillation counter (1 000–2 000 volts). Owing to their high operating voltage, they do not work efficiently into a 50 microamperemeter. Nevertheless, it is possible, by the use of current transformers, to get the maximum operating power of the trigger and ratemeter circuit down to 2 mW, which is about the same as that required by the detector and is quite acceptable. Their one disadvantage is their rather long deionization time, which results in a circuit recovery time of the order of one millisecond. This leads to considerable non-linearity of calibration at high count rates, and, in fact, makes them almost useless in scintillation counter equipment. They have, however, been used quite extensively in Geiger–Müller counter equipment.[1,2,3]

Under some conditions, it is possible to use them in scintillation counters either when the maximum count rate is low, or when it is permissible to use the mean current of the photo-multiplier as an indication of radiation intensity. In the latter case, the current from one of the electrodes of the photo-multiplier is fed into the trigger electrode circuit of the valve, which is so arranged that the valve oscillates at a frequency proportional to the mean value of the current. The oscillation frequency can then be chosen to suit the deionization time of the valve. However, this cannot be done where the instrument is required to work at

Dr. Franklin and Mr. James are at the United Kingdom Atomic Energy Research Establishment.

temperatures above about 30–40° C, because the dark current of the photo-multiplier becomes comparable with the signal current which is to be measured. In any case, considerations of the radiation energy spectrum may sometimes make it undesirable to use a mean current measurement.

The transistor would appear to combine all the advantages of hot- and cold-cathode valves. It requires no filament power, it can operate at low-h.t. supply currents and even lower h.t. supply voltages than the hot-cathode valve, and it can operate at pulse rates sufficiently high even for scintillation counters. Also, present experience indicates that it will prove at least as reliable as the cold-cathode valve. Certain difficulties arise in the application of transistors owing to the wide range of ambient temperatures over which the equipment is required to operate, i.e. approximately −40° C to +60° C. Some of the transistor characteristics change quite appreciably with temperature and these changes must be allowed for in the design of the circuits. With junction transistors it is possible to design a trigger and ratemeter circuit to operate satisfactorily over the whole temperature range with a maximum power consumption of about 1·2 milliwatts. With point-contact transistors the power consumption is appreciably greater than that of the cold-cathode valve circuit (up to 12 mW) owing to the large value of I_{co} at high temperatures.

In the design of the power supply one can either choose to supply all the power direct from batteries, even at the high voltage levels required by the Geiger–Müller or scintillation counter, or, alternatively, one can use only low-voltage batteries and transform up to the higher voltage levels after converting to alternating current. For the d.c./a.c. conversion there is the choice between vibrators, hot-cathode valve oscillators and transistor oscillators. High-voltage battery stacks have been used quite extensively in the past, and with them it has been found possible to achieve great power economy and a battery operating life not significantly different from its shelf life. However, the reliability of these batteries has left much to be desired. A good specimen stored under ideal conditions can have a life of several years, but quite frequently failure occurs after only two or three months, even under temperate conditions. In hot climates, they suffer severely from loss of moisture from the electrolyte, with resulting rapid increase of internal impedance. This weakness, added to their high cost and poor availability in many areas, has brought them into disfavour. In comparison with the vibrator, the transistor scores heavily on power economy and probably on reliability. Also, since the oscillation frequency of a transistor oscillator can be much higher than the vibrator frequency, transformers and smoothing condensers can be much smaller. In comparison with the hot-cathode valve, the transistor gives greater efficiency at lower operating voltage, and again probably scores on reliability.

We thus have a strong case for using transistors, particularly those of the junction type, both in the trigger and ratemeter circuits and in the power-supply circuits of this type of instrument.

(2) CHOICE OF TRANSISTORS

Before discussing the circuits in detail, it may be profitable to note the principal differences between point-contact and junction-type transistors and to indicate in general terms their effect on circuit performance.

The biggest disadvantage of the point-contact transistor is its much greater back current, I_{co}. Even at 20° C, the power loss due to this current is quite considerable, and in the trigger and ratemeter circuit at 60° C, it is many times more than the power taken by the whole of the remaining circuit.

The high value of I_{co} at high temperatures is particularly undesirable in cases where high stability of calibration is required. Stabilization of the supply voltage to the ratemeter circuit is then necessary, and the circuit has to be adjusted so that the stabilizer can provide the maximum power ever to be required by the circuit. The power dissipation at all temperatures therefore becomes equal to that at the highest temperature to be encountered.

Another disadvantage of the point-contact transistor is that the bottoming voltage (base-collector voltage drop when the transistor is turned hard on) is not negligible compared with the supply voltage when this is low—say 6 volts. The two main effects of this are to reduce the efficiency of the power oscillator and to increase the influence of transistor characteristics on the calibration of the ratemeter.

The last main disadvantage is the variation of current gain with circuit conditions. In the case of the junction transistor, the current gain varies very little with collector current and voltage, except at very low voltages (below 0·5 volt), and this property can be used with advantage in some circumstances. For instance, in the case of the power oscillator, it can be arranged that the current taken from the battery and the output current into the load can be made independent of battery voltage.

The lower operating frequency of the junction transistor as compared with the point-contact type, although a disadvantage, is not an important consideration in this type of application.

In addition to the disadvantages arising out of their basic characteristics, point-contact transistors of the types tested have been found to be inferior to those of the junction type in their ability to withstand the effects of temperature cycling. It is understood that certain later types of point-contact transistor are better from this point of view, but it is not known how much better they are. Tests have been carried out with a temperature cycle consisting of a rise from 25° C to 70° C in approximately 25 min and a return to 25° C in a further 35 min. This temperature range is comparable with the daily range which might be expected in a desert climate, or in a cold climate where the instruments were normally kept in a heated building when not in use.

The defects of the point-contact transistors as a result of this temperature cycling appeared as a large increase in bottoming voltage corresponding to a serious loss of gain. All important changes seemed to occur during the first 100 cycles. The limit for v_{cb} was taken as 3 volts for $i_c = 6$ mA and $i_e = 3$ mA. Of the first batch of 100 transistors, 20% failed this test during the first 100 cycles, and out of a batch of 12 transistors of later manufacture, 6 failed.

In the case of junction transistors the changes brought about by temperature cycling appear, in general, to be less severe. Small changes in gain occur, both increases and decreases, but these are unimportant because the ratemeter circuit described later is designed so that the accuracy is unaffected by changes in gain over fairly wide limits.

Of the 85 transistors tested, two showed an increase in bottoming voltage, two others an increase of I_{co} and three others developed open-circuits. These changes occurred during the first 100 temperature cycles, and the effect of further temperature cycling appeared to be to increase the magnitude of these defects rather than the number of defective transistors.

These effects require to be investigated further, but the present evidence suggests that they are likely to be much less troublesome with the junction transistor than with the point-contact type.

(3) POWER SUPPLIES FROM TRANSISTOR OSCILLATOR

(3.1) General Considerations

Two power supplies are normally required, an l.t. supply for the trigger and ratemeter circuits and an h.t. supply for the

adiation detector. The trigger and ratemeter circuits are enerally operated at about 6 volts and take a current which may ary from a few microamperes to about 200 μA when junction ansistors are used, or up to about 2 mA when point-contact ansistors are used. When the radiation detector is a Geiger–Müller counter, the h.t. supply is required to be adjustable over pproximately the range 350–450 volts, while in the case of the cintillation counter the range is approximately 1 300–2 000 volts. n both cases the total power requirement is normally less than 0 mW.

The primary power supply is generally obtained from a number f dry Leclanché torch cells. During their useful life each of these ells drops its output voltage from slightly over 1·5 volts when ew to about 1·0 volt when nearly fully discharged. To make he best use of the battery, the electronic circuits must be designed o accept this range of voltage, and this necessitates some form f voltage stabilization. The use of mercury cells would give a onger battery life and ease the stabilization problem by giving a nore constant supply voltage, during the operating life, but the act that in many parts of the world they are not easily obtainable precludes their use in geological instruments.

The detrimental effect of the high bottoming voltage (1·5–3 volts) of the point-contact transistor on the efficiency of a power oscillator has already been mentioned. In converting 5 volts to 400 volts with a power output of 10 mW, the maximum efficiency attainable was about 33%, not including losses involved in voltage stabilization.

With the junction transistor the collector-emitter voltage drop can be very much less (0·1 volt), and with a battery supply of only 3 volts and a similar output power, an efficiency of 60% can be obtained. Also the characteristics of this type of transistor are such that the output voltage can be stabilized, while retaining quite a high overall efficiency over the full range of battery voltage variation.

(3.2) Junction Transistor Oscillator

Fig. 1 shows the circuit diagram of a typical power supply for a portable Geiger–Müller counter instrument.

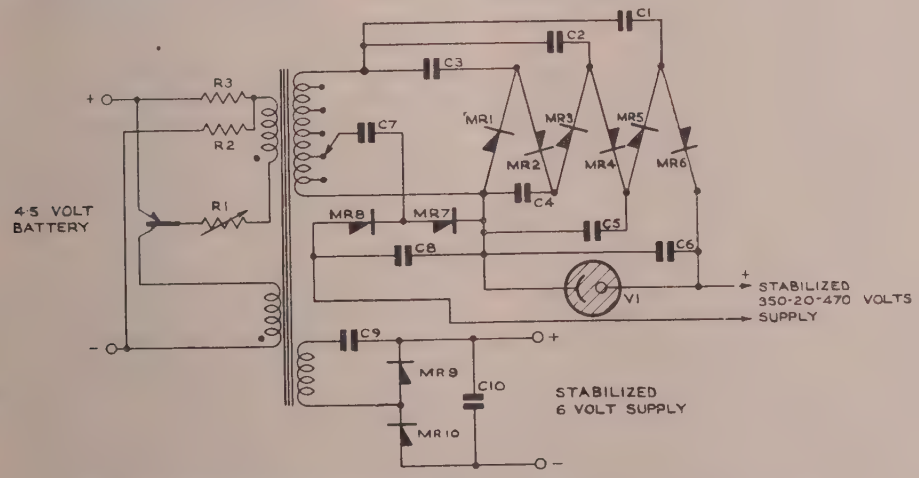


Fig. 1.—Junction transistor oscillator circuit providing stabilized h.t. and l.t. supplies.

R1	10 kΩ	C1–C6	0·05 μF	MR1–MR8	STC. type MQ8/5
R2	2·2 kΩ	C7–C8	0·02 μF	MR9–10	STC. type MDA18–1–1RG2
R3	10 kΩ	C9–C10	20 μF	V1	350 V corona stabilizer

The transistor operates as a square-wave oscillator, with positive feedback from collector to base provided via the transformer.

The mode of operation of the oscillator is closely bound up with that of the voltage stabilizing circuits. A high-ratio secondary winding of the transformer feeds into a 3-stage Cockcroft–Walton voltage multiplier and the d.c. output is stabilized by a 350-volt corona stabilizer. The through impedance of the Cockcroft–Walton multiplier is kept to a minimum by connecting all input capacitors, C_1 to C_3, direct to the a.c. input and all output capacitors, C_4 to C_6, direct back to one end of the rectifier chain, rather than using the more usual cascade connection. If this through impedance and the transformer winding impedances are made sufficiently small, the corona stabilizer fixes not only the d.c. output of the voltage multiplier but also the peak-to-peak value of the a.c. waveform across each transformer winding. As explained later the 'on' and 'off' periods of the transistor are nearly equal. Consequently the zero-to-peak square-wave voltage appearing across the base winding of the transformer is equal to approximately one-half of the peak-to-peak voltage and is therefore held constant by the corona stabilizer. The transistor base current during the 'on' period is determined by this voltage and by the sum of R_1 and R_3 and so is independent of battery voltage. The limitation of the collector square-wave amplitude by the corona stabilizer prevents the collector from bottoming, so the collector current is the product of base current and base current gain. A small fraction of this collector current is fed back to the base via the transformer to maintain the transistor conducting; the major part is shared between the load (comprising the Cockcroft–Walton multiplier and the additional voltage-doubler circuits C_7, MR_7, MR_8, C_8, and C_9, MR_9, MR_{10}, C_{10}) and the magnetizing current of the transformer. The latter is zero at the beginning of the 'on' period and rises nearly linearly until it has taken over all the current from the load and has begun to steal some of the current fed back to the base. At this point the transistor turns off and the magnetizing current in the transformer is equal to the current which flowed via the transformer into the load at the beginning of the 'on' period (neglecting the small component fed back to the base).

The voltage across the transformer now reverses and the magnetizing current flows through the load in the reverse direction, falling nearly linearly to zero (since the voltage across the transformer is again held constant by the stabilizer) when the transistor again begins to conduct.

The current in the load thus consists of triangular waves of

alternating polarity but nearly equal peak amplitude. Since no d.c. component can flow, it follows that the triangular waves must have equal areas and therefore equal durations, so the durations of the 'on' and 'off' periods must also be equal.

It will now be seen that the mean collector current (and also the oscillator output current) is fixed independently of battery voltage, and this leads to a considerable power economy with a new battery and also to very good stabilization of the transformer voltage. These currents will depend on the current gain of the transistor, and resistor R_1 is made variable so that the output current may be set up to the desired value.

At very low values of emitter current, the emitter resistance of the transistor can be so high that the loop gain around the feedback circuit is less than unity. It is therefore necessary to provide some bias current via resistor R_2, in order to make the oscillator self-starting under all conditions of loading and battery voltage. Variation of current in this bias circuit with battery voltage will to some extent upset the stabilization of the collector current. If serious, this variation—and also the inevitable loss of power in the bias circuit itself—can be avoided by use of a starting switch which introduces the bias current only during starting.

(3.3) Voltage Stabilization

As mentioned in Section 3.2, the direct voltage across the 3-stage Cockcroft–Walton voltage multiplier, and also the peak-to-peak alternating voltage across each winding of the transformer, is fixed by the corona stabilizer, V_1.

The Geiger–Müller counter operating voltage is obtained from a series combination of the multiplier and the additional voltage doubler C_7, MR_7, MR_8, C_8. The output voltage of the latter, which also is stabilized, is variable in steps of 20 volts over a range of 120 volts by selection of taps on the transformer.

A 6-volt stabilized d.c. supply for the trigger and ratemeter circuits is obtained by rectification of the output from a suitable low-voltage winding on the transformer by means of the circuit C_9, MR_9, MR_{10}, C_{10}. A stabilized supply at this voltage would be difficult to obtain by other methods.

Various systems of stabilization of transistor oscillators have been described using reference batteries for voltage standardization,[12] but it was thought highly undesirable to employ a reference battery in an instrument of this type.

When the power-supply circuit is to be used to supply a scintillation counter, the voltage per section of the Cockcroft–Walton multiplier is chosen to be equal to the required voltage between adjacent stages of the photo-multiplier. (A section is defined here as a complete voltage doubler, such as C_3, MR_1, MR_2, C_4 in Fig. 1.) Thus for an 11-stage photo-multiplier 12 sections are used and the corona stabilizer is connected across a suitable number of sections, depending upon its operating voltage. The sections chosen are those at the anode end of the chain, where the load current variations are greatest. (A resistor chain would be smaller in size than the voltage multiplier, but its power drain would be many times the total power taken by all the remaining circuits, and since high-stability resistors of several megohms would be needed, the cost also would be greater.) Adjustment of supply voltage to the photo-multiplier is achieved by supplying some or all of the sections of the voltage multiplier which are not connected direct to the stabilizer from alternative taps on the oscillator transformer.

It is sometimes more convenient to use a glow-discharge stabilizer instead of a corona stabilizer. The minimum stable current of the former may be one milliampere or more, leading to excessive power consumption if the stabilizer is used normally. Power consumption is much reduced if the stabilizer is used as a relaxation oscillator shunted across one or more sections of the voltage multiplier, as shown in Fig. 2. During the short

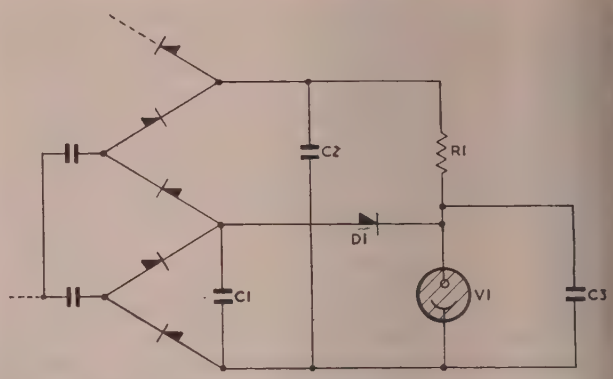


Fig. 2.—Low-current glow-discharge stabilizer.

C1, C2	0·05 μF
C3	200 μμF
R1	20 MΩ
D1	MQ3/1

conduction periods, the voltage across the stabilizer drops to about the normal burning voltage, and the selenium rectifier D_1 repeatedly discharges C_1 to this voltage. (If the current in D_1 is too small, residual ionization causes the voltage to be appreciably below the normal burning voltage, so this current should not be allowed to fall below a few microamperes.) If the mean current through D_1 is progressively increased, the result is to increase the length of the conducting periods, without substantially altering the length of the non-conduction periods. Eventually a state of continuous conduction is reached without any discontinuous changes in the mode of operation of the circuit or of the stabilized voltage. The circuit can therefore handle a very large range of input currents.

(3.4) Performance of the Power Supply

To obtain high efficiency, the circuit is so designed that with the lowest battery voltage and the worst transistor, the collector-emitter voltage does not quite fall to the bottoming voltage. With higher battery voltages, the voltage across the transistor will be higher and the efficiency consequently lower. The overall efficiency obtained on full load of 7 mW with a battery run down to 3 volts is 40%, but with a new battery giving 4·5 volts it decreases to 25%. This takes account of all losses, including those involved in stabilization, and it allows for a minimum current in the stabilizer of 2 μA when the battery voltage is minimum and the load current at its highest value.

The stabilization of the 6-volt supply is particularly good over the range of battery voltage 3 to 4·5 volts, being of the order of ±0·25%.

The load current taken from the 6-volt supply by the trigger and ratemeter circuit can vary over the range 25–200 μA, taking into account temperature effects on the standing current of the trigger-circuit transistor and the normal variations of trigger-circuit current with Geiger–Müller tube count rate. The resultant variation of the output voltage is within a range of ±1·5%. Germanium junction rectifiers would give even better regulation, but their use was avoided owing to the rapid rise of their reverse current with temperature. A variation of the corona stabilizer current from 2 to 20 μA, which is the normal operating range, due to other causes, gives a variation of output voltage which is within the range ±1%.

The power supply operates satisfactorily over a temperature range of −20° C to +60° C, although the rise of I_{co} reduces the efficiency slightly near the upper end of the range. The stabilization of the 6-volt output also is slightly affected, the output at 50° C being 2·5% higher than at 20° C and that at 60° C being

5% higher. The effect of these variations on the instrument calibration is less than might be expected, because temperature effects in the ratemeter circuit tend to compensate to some extent.

(4) TRIGGER AND RATEMETER CIRCUITS

(4.1) General Considerations

Trigger and ratemeter circuits are required to be economical in power consumption and simple in form. The triggering sensitivity must be high and also fairly stable with temperature variation and time, and the circuit dead-time after triggering sufficiently small, compared with the mean time interval between counts, to give a ratemeter calibration which does not deviate from linearity by more than a few percent. The calibration is frequently required to be constant under all conditions of use to an accuracy within about ±2%, although this requirement can sometimes be relaxed.

Point-contact transistors were used in the earlier circuits which were developed, and these circuits are of interest in that they show the relative advantages of the two types of transistor.

(4.2) Point-Contact Transistor Circuits

Fig. 3 shows a trigger and ratemeter circuit[5] using a point-contact transistor. Transformer feedback is again used as in

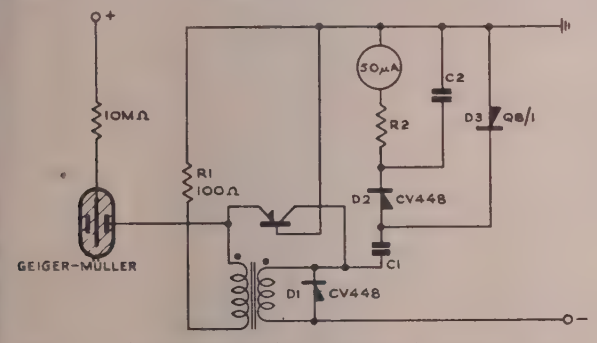


Fig. 3.—Point-contact transistor trigger and ratemeter circuit.

the power oscillator but, in this case, to the emitter. This arrangement is used in preference to a monostable circuit using two transistors,[7] in the interests of economy. When the circuit is quiescent, the emitter current is very low and the emitter input resistance is consequently high enough to keep the loop gain well below unity[4] and so the circuit is prevented from self-oscillation. If a sufficiently large positive pulse of current is passed through the emitter to the base circuit, r_e falls to a value low enough to allow the loop gain to exceed unity, and the transistor turns on with the emitter current limited by R_1. The transistor conducts for a period terminated by the rise of magnetizing current in the transformer, the energy of which is subsequently dissipated in the rectifier D_1.

The pulse duration varies with transistor current gain, so to prevent such variations from affecting the meter reading, the meter current is fed by the diode pump D_2, D_3 and C_1. C_1 is charged through D_3 between pulses and discharges into the meter circuit through D_2 during each pulse. If the pulse lasts long enough, the charge fed to the meter is independent of pulse length, but is still dependent on amplitude.

The pulse amplitude is equal to the supply voltage minus the transistor voltage drop, and in a point-contact transistor the latter is about 1·5 volts and may vary. Thus, since the supply voltage cannot easily be made large compared with 1·5 volts, the calibration may vary somewhat. A further difficulty is due to I_{co}, which varies with different transistors over the range 0·18–0·6 mA at 20° C and 0·6–2 mA at 60° C. This causes considerable power drain from the stabilized supply and also causes an appreciable and varying voltage drop across the internal base resistance of the transistor. This produces a varying forward bias at the emitter and consequently a varying trigger sensitivity, with the possibility of oscillation at high temperatures.

The first two difficulties have been largely overcome by feeding the transistor from an unstabilized supply and using the stabilized supply merely to limit the amplitude of the pulse fed to the diode

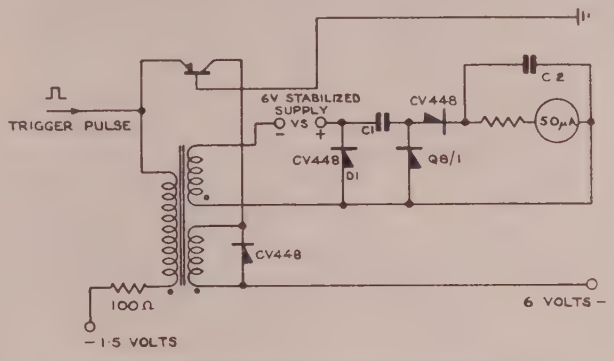


Fig. 4.—Improved point-contact transistor ratemeter.

pump, as shown in Fig. 4. The variation of trigger sensitivity with temperature does, however, remain.

The design difficulties involved in the use of point-contact transistors arise almost entirely from the large values of I_{co} and bottoming voltage, and are largely avoided by the use of junction transistors, in which both these characteristics have very much smaller values.

(4.3) Junction Transistor Circuits

The junction circuit of Fig. 5 is similar to the point-contact circuit of Fig. 3, except that the emitter, instead of the base, is

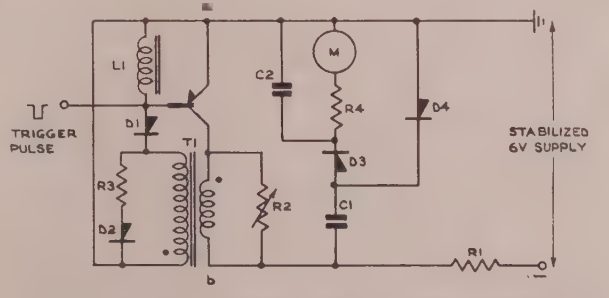


Fig. 5.—Junction transistor trigger and ratemeter circuit.

R1	2·2 kΩ	D1	CV103
R2	2·5 kΩ (maximum).	D2	CV448
R3	1·5 kΩ	D3	CV448
M	0–50 μA microammeter.	D4	MQ8/1T

The values of C1, C2 and R4 are fixed by the range of count rate.

earthed. This change is desirable because, when the transistor is in the heavily conducting state, the voltage drop between emitter and collector is smaller than that between either electrode and the base.

The collector-base current, I_{co}, of a junction transistor with the emitter open-circuited is very small, being normally less than 10 μA at 20° C, although it rises with temperature faster than that of a point-contact transistor, about doubling for each 8° C temperature rise. If the supply voltage is applied between collector and emitter, and the base is open-circuited,

the collector current, I'_{co}, is greater than I_{co} by a ratio equal to α', the current gain of the transistor as measured from base to collector. Since α' is always over 10 and sometimes over 50, the standing current of the circuit would be as high with the latter connection as for the point-contact transistor with earthed-base connection (Figs. 3 and 4).

In the circuit shown, both the emitter and the base are earthed for direct current by L_1. This arrangement has been found to give values of I''_{co} only about 1·5 times greater than I_{co} with transistors having current gains in the range 10 to 30. This means that the standing current is of the order of 50 μA at 50° C and 100 μA at 60° C. The inductance of L_1 is high compared with that of the transformer and therefore has only a negligible effect on the timing of the circuit.

The low value of I''_{co} allows the circuit to be used with the type of low-current stabilized supply described earlier and also permits a better arrangement of the metering circuit. In Fig. 3, C_1 was connected direct to the collector of the transistor, and this had the disadvantage of reducing the triggering sensitivity by an amount depending on the value of C_1. In the junction-transistor circuit, the low value of standing collector current allows of the insertion of the 2·2-kilohm resistor, R_1, and the diode pump condenser C_1 is connected to the junction between this resistor and the collector winding. This eliminates the effect of C_1 on the triggering action.

In the 'off' condition C_1 is charged to the voltage of the stabilized supply less the I''_{co} drop in R_1, which is negligible. In the 'on' condition the transistor initially draws a large current and rapidly discharges C_1. The timing of the circuit is arranged so that the average transistor remains conducting (eventually with a smaller current drawn through R_1) for about three times as long as is necessary to discharge C_1. The voltage change across the capacitor is therefore independent of the timing and consequently also of transistor characteristics over a wide range. The voltage between *a* and *b* after C_1 has discharged is only about 0·25 volt, of which approximately 0·1 volt is across the transistor. This is so low that any changes due to variations in transistor characteristics have a negligible effect on the voltage swing across the condenser.

The value of R_1 is determined by the following considerations. It is desirable to make this resistance large in order to secure a low voltage across *ab* in the 'on' condition and also to limit the flow of collector current after C_1 has discharged, in order to keep the power requirements of the circuit small. On the other hand, the size of R_1 is limited by the maximum voltage drop due to I''_{co} which can be tolerated and by the requirement that the time-constant C_1R_1 shall be short compared with the mean time between pulses. With the chosen value of 2200 ohms the maximum voltage drop at 60° C due to I''_{co} is approximately 0·2 volt, corresponding to a drop in meter reading of approximately 4%. In practice this drop tends to be corrected by a corresponding rise in the voltage of the stabilized supply, as mentioned earlier. The meter current, I, is equal to VC_1n, where V is the voltage swing across C_1 and n is the mean number of pulses per second. C_1, therefore, is equal to I/Vn and the time-constant C_1R_1 to IR_1/Vn. If R_1 is 2200 ohms and V is 5·5 volts, at full-scale meter deflection where $I = 50\,\mu$A, the value of this expression is $1/50n$, i.e. the time-constant is 1/50th of the average time between pulses.

The calibration of the circuit is affected to some extent by the finite resistances of the rectifiers D_3 and D_4. In order to avoid excessive loss of charge from C_2 during the quiescent periods it is desirable that either D_3 or D_4 shall be of a type having a high back-resistance. However D_3 takes the heavy discharge current of C_1, and, to avoid excessively long dead-time, it is desirable that this rectifier shall have a low forward resistance. D_3 is therefore a germanium rectifier and D_4 a selenium type having a low reverse current. If a sufficiently long time were allowed for charge and discharge, the final voltages across the rectifiers at the ends of the charge and discharge periods would be zero. However, owing to the very high forward resistance at low voltages, the rate of fall below about 0·2 volt for D_3 and 0·3 volt for D_4 is extremely low, and these are the approximate values at the end of the discharge and charge periods, respectively. The remanent voltage across D_3 makes the meter reading very slightly dependent on the 'on' time of the transistor, while that across D_4 causes a slight increase in the non-linearity of the meter scale. The former voltage decreases with increasing temperature, but the effect of this on the meter reading is largely compensated by the increased reverse current during the recharging of C_1. The overall effect of change of temperature of D_3 and D_4 on the meter reading is less than $\pm 1\%$ over the temperature range $-40°$ C to $+60°$ C.

In choosing the transformer ratio, two conflicting requirements must be met. During the early stages of the action of triggering, the circuit is operating as a feedback linear amplifier, and the requirement for maximum sensitivity is maximum power feedback from the collector circuit to the base circuit. During this period both circuits are of high impedance and the requirement is best met by a low-ratio transformer. However, when the transistor has been turned on, the problem is to drive sufficient current through the base circuit to keep the transistor bottomed, while the voltage across the collector winding of the transformer is falling progressively to a final value of a small fraction of a volt. This requires a higher-ratio transformer for optimum conditions, although the ratio must obviously be less than the value of α'. Taking only the latter consideration into account, the ratio might with advantage be made as high as 1 : 6, the higher-voltage winding being in the base circuit. However, this is too high for good trigger sensitivity and a satisfactory compromise is obtained with a ratio of 1 : 2.

As in the point-contact transistor circuit, a rectifier D_2 is used to prevent self-oscillation of the circuit owing to ringing of the transformer when the transistor turns off, resistor R_3 serving to limit the damping action, in order to keep the circuit dead-time short.

The dead-time of the circuit is an important consideration if a linear calibration of the ratemeter is required. Where a non-linear scale is required this can sometimes be achieved by making the dead-time of the circuit comparable with the mean time interval between successive counts. However, it is not generally advisable to do this, since the dead-time is liable to vary and cause a change in the non-linearity. In order to allow for tolerances and variations in transistor characteristics, the 'on' time of the circuit with the average transistor is made about three times that required to discharge the capacitor C_1. This, together with a subsequent recovery period, during which the energy stored in the magnetic field of the transformer is given up to D_2, comprises the circuit dead-time. It has been found possible to limit this total dead-time to less than 5% of the mean time between counts, at the count rates required for full-scale deflection of the meter. In cases where several calibration ranges are obtained by switching in different values of C_1, it is necessary, in order to preserve the proportionately short dead-time, to switch in different transformers or, in some cases, to switch to different tappings on the same transformer. The inductances of the transformers used in providing a 50 μA meter current with count rates between 10 and 1000 counts/sec vary between 1 H and 5 mH, as measured across the collector winding.

The power consumption of the circuit is made up of a meter current of 50 μA, a further 50 μA taken by the transistor in

aying on after C_1 has been discharged and a maximum of, say, 00 μA due to I''_{co}, thus making a total of 1·2 mW for a 6-volt upply.

The variation of triggering sensitivity of this circuit with temerature is of a different character from that occurring in the oint-contact transistor circuit. The r_e/i_e characteristic of a oint-contact transistor is in itself fairly stable with temperature or currents of a few microamperes or more, provided that i_e is ept constant, but this is not so in the case of the junction ransistor. The emitter resistance is given by the formula $_e = kT/e(i_e + I_0)$, where k is Boltzmann's constant, T the bsolute temperature, e the electronic charge and I_o the reverse aturation current. For the point-contact transistor I_o is small, owing to the small contact area, and the expression is therefore pproximately equal to kT/ei_e, which does not change very much ver the temperature range considered. However, I_o for the unction transistor is comparable with i_e when the latter is a few nicroamperes, and therefore, since I_o changes rapidly with emperature (being approximately doubled with every 8° C rise), he effect of temperature on the value of r_e at low currents is uite considerable. Further changes in r_e are brought about in he circuit of Fig. 5 by the passage through it of I_{co}, which also changes with temperature. The r_e/i_e characteristic of a junction transistor cannot therefore be relied upon to fix the triggering level, and, instead, an additional non-linear resistance which is not appreciably affected by temperature changes is introduced into the base circuit. This takes the form of the silicon diode, D_1. An alternative method of stabilizing the triggering sensitivity would be to apply a small positive bias voltage to the base of the transistor by means of a voltage divider across a stabilized supply. However, in view of the need to keep the resistance of the base circuit to a low value, this would lead to an appreciable wastage of power.

To trigger the circuit into the 'on' condition it is necessary to bring the current through D_1 and the base-to-emitter circuits of the transistor to the critical value required for unity gain round the feedback loop. In order to do this a pulse of current must be passed through the base-to-emitter circuit for a certain time in order to charge the stray capacitances of the circuit and because of the time delay in the rise of collector current in the transistor. This means that the pulse current required is a function of pulse width for narrow pulses, and the sensitivity of the circuit can be considered in terms of charge.

In order to allow for differences between transistors an adjustable proportion of the collector current is by-passed from the feedback path by means of a variable resistor R_2. The sensitivity of the circuit is normally set at room temperature so that the circuit triggers with a pulse of 75 μA for 2 microsec, a charge of 150 $\mu\mu$C. The sensitivity of the circuit rises with temperature and at 60° C the sensitivity corresponds to a charge of approximately 50 $\mu\mu$C. The two chief difficulties in the way of increasing the sensitivity of the circuit are as follows: first, any reduction in the critical value of base current required to trigger the circuit inevitably makes the circuit more sensitive to changes of temperature; and secondly, any increase in sensitivity involves increased damping of the transformer, to prevent spurious retriggering, and this results in increased dead-time.

This sensitivity of 150 $\mu\mu$C and also its observed change with temperature are acceptable when the circuit is to be triggered from a Geiger–Müller counter, but when the detector is a scintillation counter, the sensitivity must generally be higher and more stable.

The magnitudes of the pulses from a scintillation counter are proportional to the energies of the γ-photons incident upon the crystal, and therefore vary greatly. The pulse rate of the circuit will therefore increase rapidly as its trigger sensitivity is increased, until a point is reached at which nearly all the γ-rays are being 'counted'. For a thallium-activated sodium-iodide crystal and the γ-ray spectra normally encountered in geological survey work, this point is reached at a sensitivity well below that at which the photo-multiplier noise begins to cause frequent triggering, giving a region of nearly constant count-rate extending over a range of sensitivity of about $\pm$6 dB. Satisfactory operation on this plateau with an adequate margin of safety demands a stability of triggering sensitivity within $\pm$3 dB over the temperature range $-20°$ C to $+60°$ C.

A photo-multiplier with a gain of 200 A/lm requires a triggering sensitivity of approximately 15 $\mu\mu$C to operate in the plateau region, and at a mean count-rate of 10 000/sec takes about 1·5 mW from its high-voltage supply. If a photo-multiplier of sufficient sensitivity to operate direct into a trigger circuit of 150 $\mu\mu$C sensitivity were used, the power required would rise to about 15 mW, and it is obviously more economical in power to raise the effective trigger sensitivity by the use of a transistor amplifier. The amplifier need take only about 1·5 mW and this not necessarily from a stabilized supply. Also, dependence on high tube gain would involve either special selection of photo-multipliers or the use of photo-multipliers with a large number of stages, and the latter would result in increased size of both the photo-multiplier and the power-supply circuit.

Fig. 6 shows the complete circuit for operation from a photo-multiplier. The increased stability of the trigger circuit with variation of temperature is achieved by shunting the collector winding of the transformer T_3 with a thermistor. The variable resistor, R_5, is still required, as before, to compensate for differences in triggering sensitivity between transistors, and the value of this resistor influences the action of the thermistor. However, as would be expected, it is the transistors that require a high value of R_5 which also require most compensation. The shunt resistance required to maintain constant sensitivity, with the type of transistors and diodes used, varies from 1 000 to 400 ohms at 20° C and from 300 to 200 ohms at 60° C. This requirement can be met (approximately) by using a variable resistor of 2 500 ohms and a linear-type thermistor of 1 500 ohms at 20° C and 375 ohms at 60° C.

When using a sodium-iodide crystal the current pulses through the anode circuit of the photo-multiplier rise in less than 0·1 microsec and then decay exponentially with a time-constant of about 0·25 microsec. The transformer T_1 steps up the amplitude of these current pulses, and it is desirable to obtain as high a step-up ratio as possible. However, with high ratios the effect of the self-capacitances of the windings becomes important, and a transformer ratio of 5 : 1 has been found optimum for the particular bias current and therefore base-input resistance chosen. This value of transformer ratio gives an actual current step-up ratio, in terms of peak values, of about 1 : 3. The rectifier D_1 serves as a non-linear resistor and by-passes the larger pulses which would otherwise cause undesirable ringing in the circuit. The pulses as developed across the input transformer are only approximately 1 microsec wide, and the high-frequency characteristics of the transistor are such that, with this pulsewidth the effective value of α' is only about 2.

The amplifier stage does, however, provide the required increase in triggering sensitivity of 10 times for this type of input pulse, partly by increasing the pulsewidth.

An important requirement in the design of the amplifier is low power consumption. The standing collector current is therefore arranged to be low (0·25 mA at 20° C) and the polarity of the input pulses is such as to increase it. This current is supplied direct from the battery, but a drop in battery voltage from 4·5 to 3 volts reduces the gain by only 10%, which is acceptable. It is important to minimize any variations of i_e with voltage and

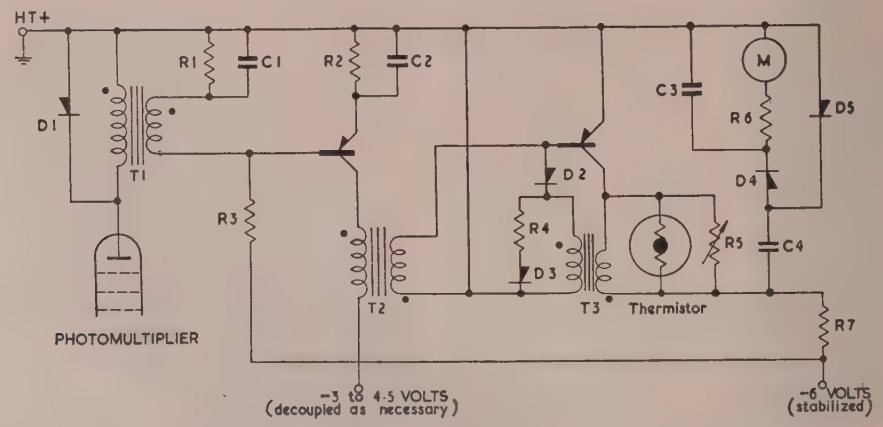


Fig. 6.—Junction transistor amplifier, trigger circuit and ratemeter as used with scintillation counter.

R1	10kΩ	D1	MQ3/1T	C1	0·05μF
R2	2·2kΩ	D2	CV103	C2	1·0μF
R3	47kΩ	D3	CV448	Thermistor. S.T. & C. type A1311/100 & A5211/100	
R4	1·5kΩ	D4	CV448		
R5	2·5kΩ	D5	MQ8/1T		
R7	2·2kΩ	M	0–50 micro-ammeter		

The values of C3, C4 and R6 are fixed by the range of count rate.

temperature, owing to the effect this has on the current gain, α', and on the input resistance.[11] This is done so far as is practicable by the use of resistors R_1, R_2 and R_3 and by supplying the bias current from a stabilized supply. With the values shown, the rise of i_e with temperature corresponds to about four times I_{co} and the voltage drop across R_2 varies from about 0·5 to 1·0 volt. The corresponding change of gain is negligible.

The output transformer, which has a ratio of 1 : 1, introduces the necessary phase reversal of the pulses and provides a d.c. connection between base and emitter of the trigger-circuit transistor in order to keep I''_{co} small.

The maximum operating frequency is limited by the operation of the trigger circuit and is of the order of 50 000 counts/sec.

(5) ACKNOWLEDGMENTS

The authors are indebted to many members of the staff of the Atomic Energy Research Establishment, including Mr. E. H. Cooke-Yarborough and Dr. J. H. Stephen, who participated in the early work on point-contact transistors, Dr. D. Taylor, Mr. E. H. Cooke-Yarborough and Dr. G. B. B. Chaplin, who offered helpful comments and criticisms on the presentation of the paper, and Mr. W. C. T. Munnoch who participated in the work with scintillation counters.

(6) BIBLIOGRAPHY

(1) Franklin, E., and Loosemore, W. R.: 'A Survey Equipment using Low-Voltage Halogen-Quenched Geiger–Müller Counters', *Proceedings I.E.E.*, Paper No. 1031 M, August, 1950 (**98**, Part II, p. 237).

(2) Franklin, E., and Hardwick, J.: 'The Design of Portable Gamma- and Beta-Radiation Measuring Instruments', *Journal of the British Institution of Radio Engineers*, 1951, **11**, No. 10.

(3) Franklin, E.: 'Cold-Cathode Valve Ratemeter having a Non-Linear Output', British Patent Application No. 21764: 1952.

(4) Cooke-Yarborough, E. H., Florida, C. D., and Stephen, J. H.: 'The Measurement of the Small Signal Characteristics of Transistors', *Proceedings I.E.E.*, Paper No. 1614 R, February, 1954 (**101**, Part III, p. 288).

(5) Cooke-Yarborough, E. H., Stephen, J. H., and James, J. B.: 'Improvements in or Relating to Transistor Circuits', British Patent Specification No. 20668: 1953.

(6) James, J. B.: 'Improvements in or Relating to Trigger Circuits', British Patent Application No. 4297: 1955.

(7) Pearlman, Alan R.: 'Transistor Power Supply for Geiger Counters', *Electronics*, August, 1954.

(8) Ryder, R. M., and Settner, W. R.: 'Transistor Reliability Studies', *Proceedings of the Institute of Radio Engineers*, 1954, p. 414.

(9) Shea, R. F.: 'Principles of Transistor Circuits' (Wiley, New York, 1953).

(10) Shea, R. F.: 'Transistor Operation: Stabilization of Operating Points', *Proceedings of the Institute of Radio Engineers*, 1952, p. 1435.

(11) Webster, W. M.: 'On the Variation of Junction Transistor Current Amplification Factor with Emitter Current', *Proceedings of the Institute of Radio Engineers*, 1954, p. 914.

(12) Light, L. H., and Hooker, P. M.: 'Transistor D.C. Convertors', *Proceedings I.E.E.*, Paper No. 1862 R, April, 1955 (**102** B, p. 775).

[The discussion on the above paper will be found on page 516.]

1.374.32 : 621.314.7

Paper No. 2064 R
Mar. 1956

A POINT-CONTACT TRANSISTOR SCALING CIRCUIT WITH 0·4 MICROSEC RESOLUTION

By G. B. B. CHAPLIN, M.Sc., Ph.D., Graduate.

(*The paper was first received* 16*th January, and in revised form* 4*th February*, 1956. *It was published in March*, 1956, *and was read before a Joint Meeting of the* MEASUREMENT AND CONTROL SECTION *and the* RADIO AND TELECOMMUNICATION SECTION, 27*th March*, 1956.)

SUMMARY

There is a wide choice of scaling devices which will operate at aximum counting rates up to several hundred kilocycles per second, ut for counting rates in the region of megacycles per second, the oice is limited almost entirely to special thermionic scaling devices r circuits using thermionic valves. Such circuits tend to be rather mplex and have a relatively high power consumption.

The paper describes some scaling circuits using transistors which ill resolve 0·4 microsec and hence count at a maximum rate of ·5 Mc/s. The transistors are the normal point-contact type, and the ircuits are simple, they have wide tolerances and are economical in ower consumption.

Features which contribute to the short resolving time are the pre- ention of bottoming of collector potential and the absence of apacitors. A typical scale-of-10 circuit uses seven transistors, seven ulse transformers and 14 crystal diodes.

LIST OF SYMBOLS

i_e, i_b, i_c = Instantaneous current flowing in the emitter, base and collector electrodes, respectively.

I_e, I_b, I_c = Current supplies to the emitter, base and collector, respectively.

r_e, r_b = Internal emitter and base resistances, respectively.

R_e, R_b = External emitter and base resistances, respectively.

v_e, v_b, v_c = Instantaneous potentials of the emitter, base and collector electrodes, respectively.

α = Ratio of i_c to i_e above the knee of the i_c/v_c curve for the particular value of i_e being considered.

i_{c0} = Collector-to-base leakage current for $i_e = 0$ and a specified value of v_c.

(1) INTRODUCTION

The function of a scale-of-*n* circuit is to accept input pulses hich may be random in their occurrence, and to emit one utput pulse for every *n* input pulses. An important criterion f any scaling circuit is its resolving time, which is the minimum eparation between adjacent input pulses that the circuit can ccept without counting them as a single pulse.

A resolving time of about 250 microsec, corresponding to a ounting rate of 4 kc/s, can conveniently be attained by cold-athode-tube scalers, and rates up to several hundred kilocycles er second can be obtained by magnetic devices or junction ransistors, but for rates in the region of megacycles per second, he thermionic valve has had no serious competitor. Such ircuits either use special scaling valves[2,3] or tend to be rather omplex and have a relatively high power consumption; a ypical scaler uses two double triodes and several diodes per inary stage.[4]

By preventing bottoming of collector-to-base potential[5,6,7] nd eliminating coupling capacitors, a point-contact transistor caler has been developed which will resolve 0·4 microsec and hence count at a maximum rate of 2·5 Mc/s. The circuit is simple and consumes little power.

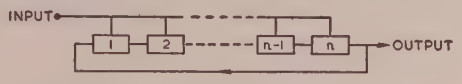


Fig. 1.—Block schematic of ring scaling circuit.

Fig. 1 is a block schematic of *n* bistable circuits connected in a ring, such that the following rules are obeyed:

One, and only one, stage is in the 'on' state during steady-state conditions.

An input pulse causes the 'on' state to move to the next stage in the ring.

Thus *n* input pulses cause the 'on' state to be cycled once around the ring, whereupon an output pulse is produced.

(2) BASIC CIRCUIT

Fig. 2 shows a scale-of-3 circuit which embodies these rules. Three point-contact transistors have their emitters joined together and connected to a common current supply I_e. So far as direct currents are concerned, each base is connected to earth by a resistor R_b, and the collectors are taken to a common negative potential. If the resistors R_b have resistance greater than a certain minimum value, and if a current greater than a certain minimum value flows into any emitter, the emitter impedance will be negative. Under these conditions it is impossible for the current, I_e, to be shared by two or more transistors, for if two or more were conducting they would be in unstable equilibrium. Thus the current, I_e, ensures that at least one transistor is in the 'on' state, and the negative impedance at the emitters ensures that only one is in the 'on' state.

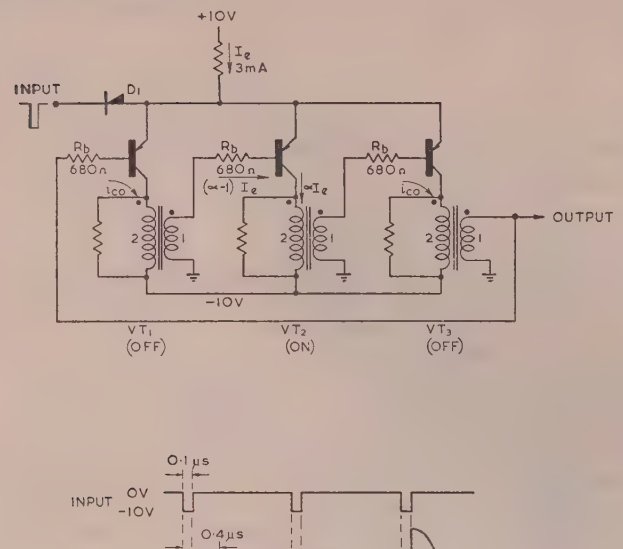
If it is assumed that one of the transistors in Fig. 2 is 'on', e.g. VT_2, its emitter current, I_e, produces a collector current of αI_e and hence a base current of $(\alpha - 1)I_e$. This results in a negative base potential of $(\alpha - 1)I_eR_b$. The base potentials of the other transistors will be negative to earth by an amount $i_{c0}R_b$, since zero emitter current is assumed, and if the potential of the base of VT_2 is sufficiently negative to cut off the other transistors by holding the potential of the emitters below that of any of the other bases, the state is stable.

So far the circuit obeys the first of the two rules stated in Section 1, but to obey the second rule it is necessary to include a reactive component between stages.

This is necessary because it is impossible for a given stage simultaneously to change its state to that of the preceding stage and indicate its state to the following stage. A memory of the initial state must therefore be retained during the change-over, and this can be achieved conveniently by including an inductor in series with each collector.

If a sufficiently large negative input pulse is applied to the cathode of D_1, the potential of the common emitters will be taken more negative than that of any of the bases, and the

Dr. Chaplin is at the United Kingdom Atomic Energy Research Establishment.



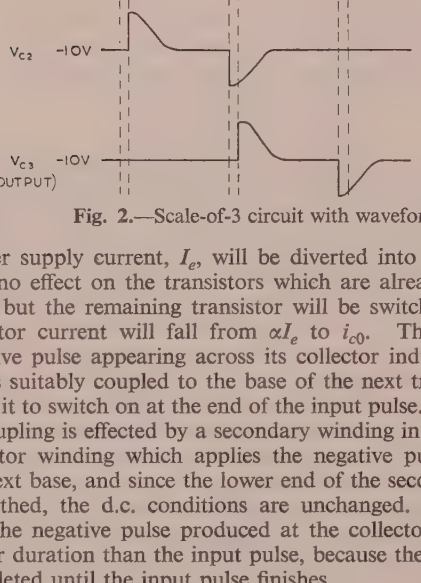


Fig. 2.—Scale-of-3 circuit with waveforms.

emitter supply current, I_e, will be diverted into D_1. This will have no effect on the transistors which are already in the 'off' state, but the remaining transistor will be switched off and its collector current will fall from αI_e to i_{c0}. This results in a negative pulse appearing across its collector inductance, and if this is suitably coupled to the base of the next transistor it will cause it to switch on at the end of the input pulse.

Coupling is effected by a secondary winding in phase with the collector winding which applies the negative pulse directly to the next base, and since the lower end of the secondary winding is earthed, the d.c. conditions are unchanged. It is essential that the negative pulse produced at the collector should be of longer duration than the input pulse, because the transfer is not completed until the input pulse finishes.

One of the main features of the circuit is that the potential difference between collector and base is prevented from falling below a certain minimum value. This corresponds to the knee of the i_c/v_c characteristic, and thus bottoming of the transistor, and hence excessive carrier storage, is prevented, which enables it to be switched off rapidly.[5,6,7]

In fact, the input pulse need be of no greater duration than 0·1 microsec for the type of point-contact transistor which has been available in this country for several years. The duration of the collector pulses need only be large compared with the input pulses, and so 0·4 microsec is adequate; this requires a collector inductance of about 0·25 mH.

(2.1) D.C. Conditions

Before assigning values to I_e and R_b it is necessary to determine the minimum voltage by which the base of the conducting transistor must be negative with respect to that of any cut-off transistor. This minimum voltage is the difference between the emitter-to-base voltage of the conducting transistor and the emitter-to-base voltage necessary to cut off emitter current in any of the other transistors.

Measurements made on type EW51 transistors show that for an emitter current of 3 mA and a collector potential of −10 volts the most positive emitter-to-base voltage likely to be encountered is +0·5 volt, an average value being 0·35 volt. If I_e is now reduced to zero the emitter assumes a potential not more negative than −0·2 volt. This is the voltage required to cut off the emitter current completely, but in fact it is only necessary to reduce the emitter current to a value which ensures a large enough positive resistance between emitter and earth in the circuit of Fig. 2. This resistance is $r_e + (1 - \alpha)(r_b + R_b)$, and so is positive if r_e exceeds a value two or three times that of $(r_b + R_b)$. The emitter resistance, r_e, is nearly inversely proportional to emitter current[8] and is greater than 10 kilohms at an emitter current of 2 μA. The emitter of a cut-off transistor may thus safely pass this current, which reduces the cut-off bias to zero voltage.

Under these conditions the total maximum cut-off bias required is only 0·5 volt, and so the criterion for stability in the circuit of Fig. 2 is

$$[(\alpha - 1)I_e + i_{c0\,min} - i_{c0\,max}]R_b > 0{\cdot}5 \text{ volt}$$

This condition is satisfied by a wide range of values. For example, if $I_e = 3$ mA and $R_b = 680$ ohms, if α is given its minimum value of 2, and if the extremes of i_{c0} (i.e. $i_{c0\,max} - i_{c0\,min}$) are given the maximum value of 2 mA, the left-hand side becomes 0·68 volt, which allows an adequate margin of safety.

Similar tests on type 3X/100N transistors, for emitter currents of 3 mA, 2 μA and zero, yielded extreme emitter potentials of +0·4, −0·15, and −1·5 volts, respectively. If the case of zero emitter current is not considered, these result in a total cut-off potential of 0·55 volt in the right-hand side of the inequality. Although this is slightly worse than for the type EW51 transistor it is compensated in the left-hand side of the inequality by the fact that there is less spread in i_{c0}, the maximum value of $(i_{c0\,max} - i_{c0\,min})$ being only 1 mA.

The emitter current must be large enough to produce the stable states as already mentioned, and also to provide sufficient collector current to operate the transformer memories. The maximum allowable emitter current is set either by the collector bottoming, owing to voltage drop in R_b or collector power dissipation. Bottoming occurs when $I_e \simeq v_c/(\alpha - 1)R_b$, which becomes 5 mA for a current gain of 4 and the values shown in Fig. 2.

(2.2) Input Pulse Requirements

The minimum amplitude of input pulse required depends on the highest current gain encountered. For a maximum current gain of 4, the lowest base potential is −6·1 volts, and so an input pulse must reach at least −7 volts. The top level of the input pulse need be no higher than earth potential since the forward conducting voltage of D_1 is sufficient to allow the common emitters to rise above the most positive base potential.

It is often required to cascade several scalers, and a convenient negative pulse for driving the next scaler is available on any base winding, although amplification may be necessary if the next scaler requires an excessive input amplitude. The transformer inductances in the next scaler should be doubled to ensure that the collector pulses are wider than the input pulse which is now the 0·4 microsec pulse from the first scaler. Alternatively, the input pulse can be given a positive bias so that only the negative tip is used, effectively narrowing it and so eliminating the need for increasing the inductance.

(2.3) Dynamic Stability

The static stability criterion derived in Section 2.1 is a necessary, t not sufficient, condition for the states of the scaler to be ible. Even if all but one of the transistors in the scaler are moved, and the emitter current is thus forced to flow into the maining transistor, it is possible for instability to occur.

Consider the circuit of Fig. 3. The transistor has a defined

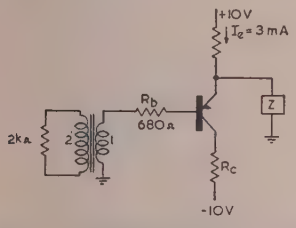


Fig. 3.—Dynamic-stability test circuit.

mitter current and resistive loads both in the base and collector. n addition, there is some impedance between emitter and earth rranged in such a way that it does not affect the d.c. conditions. f the collector load, R_c, is sufficiently large to produce adequate ottoming of the collector potential, the circuit will be stable or any positive value of Z. If R_c is now reduced to prevent ottoming, the circuit will be stable only if Z is sufficiently large.[6]

Batches of transistors of two different types tested in this ircuit for the minimum value of Z for stability gave a spread rom 300 ohms to 700 ohms when Z was resistive.

If Z is capacitive the spread in threshold value should be con-iderably greater, since it also depends on the frequency response f the transistor. This was confirmed by experiment, the maxi-num capacitance for stability ranging from 300 $\mu\mu$F to 33 $\mu\mu$F.

If the ring scaler of Fig. 2 is considered it can be seen that ach emitter has the other emitters in parallel with it. It is herefore essential that, in, for example, a ring-of-10 circuit, the esulting impedance of nine non-conducting emitters in parallel hould have a resistive component greater than 700 ohms and a apacitance less than 33 $\mu\mu$F. The first requirement is easily atisfied, since the incremental emitter resistance under these onditions is at least 10 kilohms, which results in a total impedance f 1·1 kilohms. The second requirement is easily met as far s the transistors are concerned, since their emitter capacitance s in the region of 1–2 $\mu\mu$F, which is a total maximum capacitance f 20 $\mu\mu$F, but care must be taken with the siting of the com-noned emitters to prevent the wiring from unduly increasing he effective capacitance. In practice, it was found that, with he worst combination of transistors in such a scale-of-10 circuit, here was a safety factor of at least 10 $\mu\mu$F before instability ccurred. A further improvement in this safety factor can be btained by splitting the ten transistors into two sets of five ach, and supplying their emitter current through two diodes s shown in Fig. 4. Assuming the capacitance of a diode is imilar to that of an emitter, the total capacitance connected to

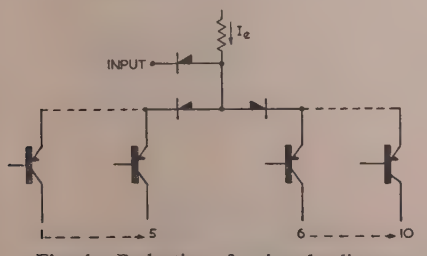


Fig. 4.—Reduction of emitter loading.

any emitter is approximately halved and the physical layout of the circuit is simplified. The circuit tolerances, however, will be worsened owing to the voltage dropped across the conducting diode.

(3) DEFINED BASE POTENTIALS

Although the circuit of Fig. 2 is economical in components and is adequate for many requirements, it can be improved in several respects.

Desirable improvements are a reduction in the required input voltage, a known collector-to-base potential in the steady state, and the ability to tolerate increased capacitance on the commoned emitters. These features can be achieved by replacing each base resistor by a current supply and a pair of catching diodes as shown in Fig. 5.

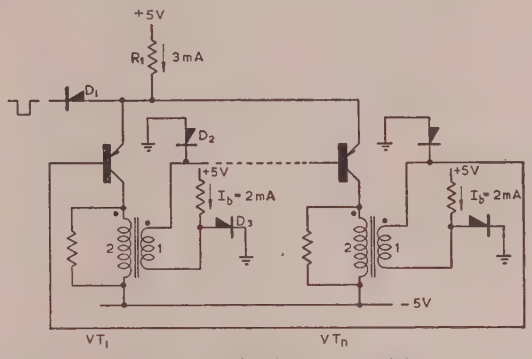


Fig. 5.—Defined base potentials.

The two diodes are used to define the upper and lower limits of the base potential. In the 'off' state, D_3 is held conducting by the excess of I_b over i_{c0}, and the base potential is near +0·5 volt. A base supply current of 2 mA is adequate for type 3X100N transistors, but it must be increased to 3 mA for type EW51 transistors. In the 'on' state D_2 conducts and defines the base potential near −0·5 volt. The change of base potential between the two states is adequate for presently available point-contact transistors if the diodes are either germanium point-contact or silicon junction types, but it can be increased if desired by returning the D_2 anode voltage to a negative bias potential.

The limited excursion of base potential in Fig. 5 enables the input pulse to be reduced in amplitude. The upper level, however, must be sufficiently positive to allow the emitters to rise above the base potential of the 'off' state, otherwise a stable state is possible with all transistors switched off, and the lower level must be sufficiently negative to cut a transistor off when its base potential is near −0·5 volt. If the input pulse is produced from another similar scaling circuit, it can be obtained from a third winding on one of the coupling transformers. It is then permissible to return one end of this secondary winding to earth potential instead of to a positive bias, because the input consists of alternate positive and negative pulses (see the waveforms of Fig. 2), and although an unwanted stable state with all transistors switched off is possible, it will be broken by the first positive pulse to arrive. Thus the pulse amplitude need not be more than 2 volts.

Another result of defining the lower limit of the base potential is that the collector-to-base voltage in the 'on' state, and hence the voltage available to establish current in the transformer, is the same for all transistors, and so the h.t. voltage can be reduced. The collector dissipation is now approximately the product of the negative h.t. voltage and αI_e.

The external base impedance in the 'on' state for a transistor in the circuit of Fig. 5 is the incremental forward impedance of the diode D_2. Since this impedance is lower than the series impedance of R_b and the transformer winding in Fig. 3, the stability of the circuit is substantially improved. The maximum allowable capacitive load on the emitter for stability is thereby increased from 33 to 60 $\mu\mu$F.

(3.1) Indication of State

A common requirement in a scaling circuit is that a voltage should be available from each stage to indicate its state. The base waveform in Fig. 5 may not have sufficient amplitude for this purpose, but a waveform can be obtained from the collector as shown in Fig. 6.

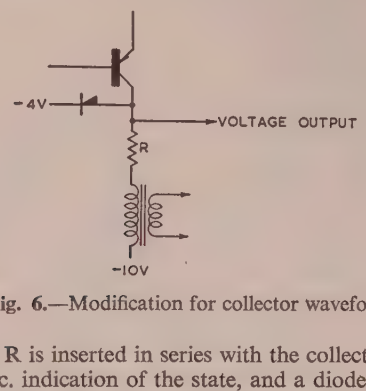


Fig. 6.—Modification for collector waveform.

A resistor R is inserted in series with the collector inductor to provide a d.c. indication of the state, and a diode is included to prevent bottoming. The value of R should be chosen so that, with the minimum current gain in the transistor, the diode just conducts in the 'on' state.

A suitable device for monitoring the collector voltage is the DM70 tuning indicator, which requires a heater current of 25 mA at 1·4 volts and an anode supply of +70 volts. If one side of the heater is earthed, and the grid is connected to the collector of the transistor, the anode will fluoresce whenever the transistor is in the 'on' state.

Alternatively the collector current can be monitored by some electro-mechanical device such as a meter.

(4) OTHER CIRCUIT ARRANGEMENTS

The number of transformers required for a scale-of-n circuit need only be $n - 1$, since one of the transistors can be given a preferred state which is selected when all the memory inductors are inoperative. This is illustrated in Fig. 7.

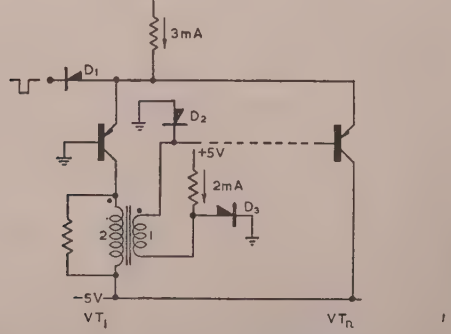


Fig. 7.—Elimination of the nth coupling circuit.

The transformer is omitted from the collector of VT_n, and the base of VT_1 is returned to earth potential. When VT_n is switched off by the input pulse its base potential rises until it is 'caught' by D_3 at about +0·5 volt. All the bases except that of VT_1 are now at a potential of +0·5 volt, which causes VT_1 to be switched on when the input pulse ceases. The potential difference between the base of VT_1 and that of any of the other transistors is now the 0·5-volt drop across a single diode. This is only half that of the circuit of Fig. 5, and it is, in fact, the minimum cut-off bias derived in Section 2, but it appears to be quite sufficient in practice. A greater factor of safety can, however, be obtained by returning the pairs of diodes to bias potentials of, say, ±1 volt.

A simple method of resetting the circuit to its initial state is to apply an input pulse of longer duration than the collector pulses, which has the effect of selecting VT_1.

If all but two stages are omitted, the circuit comprises a binary scaler which includes two transistors, three diodes and one transformer. Since there are only two stages there is no danger of dynamic instability, and so the base diodes can be placed in series with the transformer secondary as shown in Fig. 8. This arrangement allows a greater voltage swing on the transformer with a consequent small improvement in resolution.

An additional transformer can be inserted in the VT_2 collector to provide an output pulse which can be loaded without affecting the operation of the circuit.

(5) PRACTICAL SCALE-OF-10 CIRCUIT

A practical scale-of-10 circuit might consist of the binary scaler of Fig. 8 followed by five stages of the ring scaler of Fig. 5

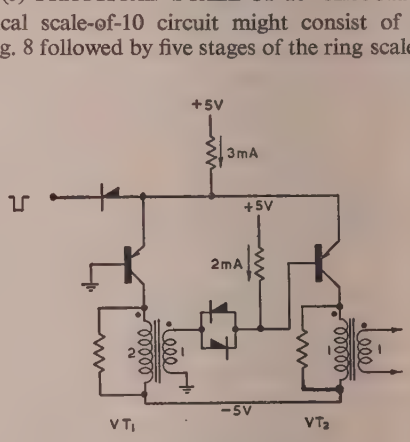


Fig. 8.—Binary scaler with modified coupling and separate output transformer.

The resolution of the complete circuit is determined by the binary scaler, which has therefore a transformer primary inductance of 250 μH resulting in a resolving time of 0·4 microsec. The scale-of-5 circuit has primary inductances of 500 μH.

The transformers have ferrite cores, and care is taken to keep the self-capacitance of the windings low. The resolution of the circuit is limited by the transformer, and more work is necessary to determine the optimum core material and geometry, and the effect of the unidirectional primary current.

The complete circuit comprises 7 transistors, 7 transformers, 14 diodes and 9 resistors, and the power consumption is about 120 mW.

(6) CONCLUSIONS

The point-contact transistor enables simple bistable circuits to be constructed having wide tolerances and an upper frequency

mit of several megacycles per second. The use of inductors nstead of condensers for reactive components allows the upper requency limit set by the transistor to be approached by the omplete circuit.

The tolerances of the circuits described were derived using the xtreme parameters of a batch of 30 transistors, but the chance f a single transistor combining all these extremes is very small, nd the chance of two or more such transistors being selected or a given ring scaler is even more remote.

In practice, no trouble has been encountered using component alues well outside the calculated tolerances of the circuits, and t is concluded that point-contact transistor circuits of this type ompare favourably with thermionic-valve circuits of similar esolution.

(7) ACKNOWLEDGMENTS

The basic ideas for the scaler were evolved at Manchester University and the development was carried out at A.E.R.E., Harwell. Another scaling circuit, which makes use of the prevention of bottoming of collector potential was developed independently by Wells,[9] but it is limited to a resolving time of 4·0 microsec because of capacitive coupling.

The author is most grateful to Mr. E. H. Cooke-Yarborough for many helpful suggestions. The help of Mr. P. Kerry, who carried out most of the experimental work, is also gratefully acknowledged.

(8) REFERENCES

(1) Cooke-Yarborough, E. H., Bradwell, J., Florida, C. D., and Howells, G. A.: 'A Pulse-Amplitude Analyser of Improved Design', *Proceedings I.E.E.*, Paper No. 933 M, March, 1950 (**97**, Part III, p. 108).

(2) 'Fast Counter Circuits with Decade Scaler Tubes', *Philips Technical Review*, 1955, **16**, p. 360.

(3) Fan, S. P.: 'The Magnetron Beam Switching Tube', *Journal of the British Institution of Radio Engineers*, 1955, **5**, p. 335.

(4) Wells, F. H.: 'A Fast Amplitude Discriminator and Scale of 10 Counting Unit for Nuclear Work', *Journal of Scientific Instruments*, 1952, **29**, p. 111.

(5) Cooke-Yarborough, E. H.: Discussion on 'A Method of Designing Transistor Trigger Circuits', *Proceedings I.E.E.*, July, 1953 (**100**, Part III, p. 245).

(6) Chaplin, G. B. B.: 'The Transistor Regenerative Amplifier as a Computer Element', *ibid.*, Paper No. 1647 R, March, 1954 (**101**, Part III, p. 298).

(7) Baker, R. H., Lebow, I. L., and McMahon, R. E.: 'Transistor Shift Registers', *Proceedings of the Institute of Radio Engineers*, 1954, **42**, p. 1152.

(8) Cooke-Yarborough, E. H., Florida, C. D., and Stephen, J. H.: 'The Measurement of the Small-Signal Characteristics of Transistors', *Proceedings I.E.E.*, Paper No. 1614 R, February, 1954 (**101**, Part III, p. 288).

(9) Wells, F. H.: 'Transistors in Scaling Circuits' (A.E.R.E. Report EL/R 1616; 1955).

[The discussion on the above paper will be found on page 516.]

621.374.32 : 621.314.7

Paper No. 2076 R
Mar. 1956

A JUNCTION-TRANSISTOR SCALING CIRCUIT WITH 2 MICROSEC RESOLUTION

By G. B. B. CHAPLIN, M.Sc., Ph.D., and A. R. OWENS, M.Sc., Graduates

(*The paper was first received* 30*th January, and in revised form* 15*th February*, 1956. *It was published in March*, 1956, *and was read before a Joint Meeting of the* MEASUREMENT AND CONTROL SECTION *and the* RADIO AND TELECOMMUNICATION SECTION, 27*th March*, 1956.)

SUMMARY

When junction transistors are used in conventional scaling circuits the maximum speed of operation is limited by the associated circuit, mainly owing to the use of capacitors, which require time to charge and discharge. The limiting speed of a transistor itself, which depends on the switch-on and switch-off times of current, is generally several times higher than this, but cannot be taken advantage of owing to the associated circuit.

The basic binary scaling circuit described in the paper overcomes this difficulty by dispensing with capacitors, a differentiating transformer being used instead for coupling.

In this way the speed of the circuit depends only on transistor characteristics. With currently-available low-frequency junction transistors ($f_{co\alpha} \simeq 500$ kc/s) the circuit is capable of reliably resolving 2 microsec.

The basic binary scaler is readily adapted to the formation of a scale-of-5 circuit using three binary stages. When this is preceded by another binary scaler, the result is a scale-of-10 circuit with the same resolving capabilities as the original binary circuit.

The circuits have wide tolerances and are insensitive to transistor variations. A complete scale-of-10 circuit uses eight transistors, ten diodes and five transformers.

(1) INTRODUCTION

The maximum operating speed of junction-transistor devices is set by the switch-on and switch-off times of collector current in the transistor, and also to a lesser extent by the effect of the current required to charge the collector-base capacitance when the collector voltage changes. If the collector is allowed to bottom during operation, i.e. the operating point enters the region of collector-current saturation, the resultant carrier storage will also decrease the possible maximum speed, since the switch-off time will be considerably lengthened. Under suitable conditions, switch-on and switch-off times of the order of less than 2 microsec are possible with low-frequency junction transistors.

A minimum resolution time of the same order is therefore indicated for a binary scaler using junction transistors. Hitherto this limit does not appear to have been attained using low-frequency transistors ($f_{co\alpha} \simeq 500$ kc/s), resolution times of the order of 10 microsec being more usual. Even with high-frequency units, results obtained have not been as good as might be expected.[1] This would appear to be a limit imposed by circuit arrangements rather than by the transistor itself.

The circuit which is described in the paper represents an attempt to achieve the maximum speed as defined above of which a junction-transistor binary scaler seems capable.

The circuit is intended as the basis of a decade scaler, the emphasis being on the retention of the maximum speed of the device in the complete scaler. The final product is therefore a junction-transistor decade scaler with a resolving time of 2 microsec.

(2) GENERAL

Conventional cross-coupled circuits use capacitive memories and coupling, and very often also have capacitors included in their gating circuits; the output indication is by means of collector voltage. The maximum operating speed of these circuits tends to be determined by the charging and discharging times of the capacitors; collector capacitance will also have a retarding effect owing to the relatively large voltage swings occurring at the collectors. The circuit which will now be described does not include any capacitors, and the collector-voltage swing is substantially reduced where this is liable to affect the operating speed of the circuit.

(2.1) Basic Binary Scaling Circuit

The circuit consists essentially of a defined current supply I_e which is applied to the emitters of two transistors connected together. Only one transistor conducts at any one time; the other remains cut off. Application of an input pulse changes the states of both transistors, so that binary scaling is possible.

The circuit may be regarded as consisting of two feedback loops combined together.

(2.1.1) D.C. Feedback Loop.

The d.c. feedback loop (Fig. 1) consists of a common emitter coupling and a collector-base coupling combined to form a bistable circuit, and it bears some resemblance to a circuit already published.[2]

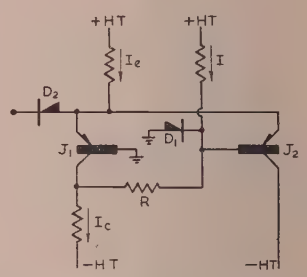


Fig. 1.—Basic bistable circuit (d.c. feedback loop).

In the first stable state with J_1 conducting, J_1 is bottomed, so that its collector potential is only 0·1 volt negative to the emitters. The voltage drop in the coupling resistor R is sufficient to take the base potential of J_2 positive with respect to that of the emitter, thus holding J_2 cut off.

In the second stable state J_1 is non-conducting, and the collector supply current I_c, now flowing in R, pulls the base potential of J_2 negative until it is caught on D_1. The base potential is thus fixed at about 0·5 volt below earth. The emitter of J_2 in this condition is 0·2 volt positive with respect to its base, and therefore 0·3 volt negative with respect to the base potential of J_1 (which is at earth potential), and so J_1 is held non-conducting.

The transition from the first state to the second is accomplished by the application of a negative pulse to the cathode of D_2. This lowers the common emitter potential below earth, and therefore below the base potential of J_1. J_1 is therefore cut off, and the collector current is diverted into R. The base potential

Dr. Chaplin and Mr. Owens are at the United Kingdom Atomic Energy Research Establishment, Electronics Division.

J_2 is then held negative to earth on the catching diode D_1. When the cathode potential of D_2 is released, and the emitter potentials are allowed to rise again, J_2 conducts first, since its base potential is more negative than that of J_1. The second state is thus set up.

(2.1.2) **A.C. Feedback Loop.**

To obtain a transition from the second back to the first stable state, a second feedback loop is necessary (Fig. 2). This is formed

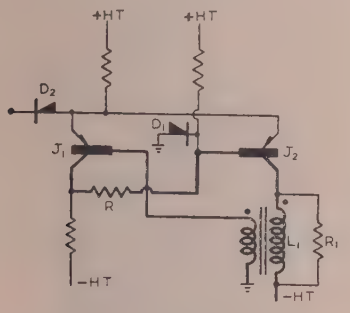


Fig. 2.—A.C. feedback loop.

by the collector-base coupling resistor R and a transformer coupling from the collector of J_2 to the base of J_1.

When the circuit is in the second stable state, with J_2 conducting, a negative pulse applied to the emitters through D_2 takes the emitter potential of J_2 negative with respect to that of its base. Since J_2 was not initially in the bottomed state, a rapid switch-off of collector current occurs in J_2. The sharp change in current causes a single negative overshoot of voltage at the collector, owing to the differentiating action of L_1, which is critically damped. This is then transformed down by the secondary winding and applied to the base of J_1, taking its potential negative. Removal of the input pulse at this point, before the transformed pulse has had time to decay, results in the emitter potentials rising whilst the base potential of J_1 is still held negative. J_1 therefore conducts first, and the first stable state is set up.

(2.2) Choice of Values

(2.2.1) **D.C. Conditions.**

When the circuit is in the first stable state, with J_1 conducting, the voltage drop in R must be sufficient to hold J_2 cut off. J_1 is in the bottomed condition with its collector at a potential of $-0{\cdot}1$ volt with respect to that of the emitters. To hold J_2 cut off, a base-emitter potential of $+0{\cdot}1$ volt is sufficient, and so a base-emitter bias of $+0{\cdot}2$ volt provides an adequate margin of safety. A potential difference of $0{\cdot}1 + 0{\cdot}2 = 0{\cdot}3$ volt is therefore required across R, i.e. $IR = 0{\cdot}3$ volt.

In the second stable state the collector potential of J_1 is determined by the voltages dropped across D_1 and R. The voltage drop in D_1 is fixed at approximately $0{\cdot}5$ volt, so that the collector potential is largely determined by R. By decreasing R and increasing I in proportion, the voltage swing at the collector may be kept small. In this way, the retarding effect of collector capacitance is substantially reduced.

With $R = 220$ ohms, I is set at $1{\cdot}5$ mA, and the resulting collector-voltage swing between the two states is only $1{\cdot}6$ volts.

By replacing R with a diode, the necessary shift in level from the collector of J_1 to the base of J_2 may be achieved by the forward conducting potential of the diode.

The voltage swing at the collector of J_1 is thereby further reduced to about $0{\cdot}8$ volt.

(2.2.2) **Transformer Design.**

The coupling transformer is required to apply a negative pulse to the base of J_1, which must satisfy the following conditions:

(*a*) It must be of such an amplitude as to hold the base potential of J_1 negative with respect to the base potential of J_2 at the instant of the removal of the input pulse.
(*b*) It must be greater in width than the input pulse.

Condition (*a*) is necessary to ensure that, when the input pulse is removed, allowing the circuit freedom to set up a stable state, J_1 becomes conducting, since its base potential is more negative than that of J_2. If this condition is not satisfied, the previous state will be retained, with J_2 conducting.

Since the base potential of J_2 is $-0{\cdot}5$ volt, owing to the drop in D_1, a transformed overshoot of at least $0{\cdot}8$ volt amplitude is required to satisfy this condition.

Condition (*b*) is also necessary to secure the transition, since, if no pulse is present in the transformer when the input is removed, the circuit has no 'memory' of its previous state. Since the second state described above is the easier for the circuit to attain (i.e. a preferred state) this will naturally be taken up, and no change will have taken place.

Condition (*b*) is satisfied by adjusting the input pulse to be narrower than the transformer overshoot. This sets a maximum limit to the input pulse width.

The width of the transformer overshoot is proportional to $\sqrt{L_1(C_s + C_c)}$, where C_c is the collector capacitance and C_s represents other stray capacitances (including coil capacitance). C_c is of the order of 50 $\mu\mu$F for the type of transistor used. If the width of the overshoot is large, the resolving time of the circuit will be governed by this width. Decreasing the width by reducing L_1 results in a progressive improvement in resolving time, until eventually transistor characteristics become the dominating factor, and no further improvement is possible. L_1 is therefore chosen to be small enough so that the transistor itself is the factor which limits the speed of operation.

The transformer ratio is chosen so that the damping of the primary winding due to the transformed base impedance of J_1 is less than the critical value, at the same time yielding a sufficient voltage swing at the secondary to satisfy condition (*a*). Extra damping is supplied by the shunting resistance R_1 to ensure that the collector is critically damped.

The final choice of L_1 is bound up with the selection of the emitter current, which in turn has a bearing on the operating speed of the circuit.

(2.2.3) **Values of Emitter Current and Inductance.**

The transistor J_2 is used as a switch interrupting a current flowing in an inductance to produce an overshoot of voltage. The amplitude of the overshoot depends on the inductance, the current which is being interrupted, and the switch-off rate of current in the transistor. The required amplitude is determined by the ratio of the transformer. If this ratio is 4 : 1 the damping imposed by the secondary loading is reduced to a tolerable value which does not greatly affect the primary winding. The overshoot required at the primary winding is therefore at least $3{\cdot}2$ volts (since $0{\cdot}8$ volt is required at the secondary winding). Use can now be made of the data contained in Fig. 3 (which is a plot of overshoot amplitude against inductance for two emitter currents, the inductance in each case being critically damped by a shunting resistor) and Fig. 4 (which is a plot of the distribution of switch-on time in a batch of over 50 transistors). The transistor used for the measurements shown in Fig. 3 exhibits average characteristics, corresponding to the peaks in Fig. 4.

For a transistor operating in the linear region the switch-on and switch-off times of current both depend on the same sets of transistor parameters.[3] From Fig. 4 it may be deduced that

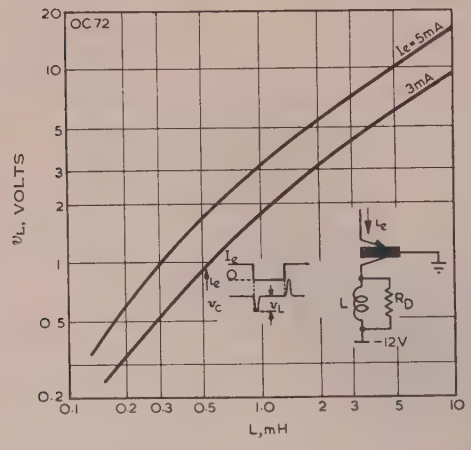


Fig. 3.—Collector overshoot voltage v_L as function of inductance.
The coils are wound on LA2 ferrite core.

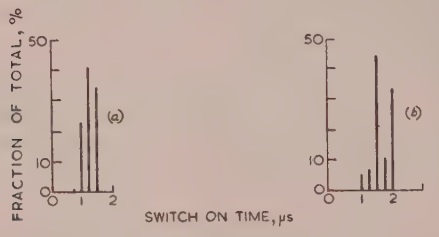


Fig. 4.—Distribution of switch-on times in a batch of transistors type OC72.
$I_e = 5$ mA.
(a) $I_c = 2$ mA (b) $I_c = 3$ mA.

increases in switch-on times of up to 30% above the average value may be encountered in a batch of transistors, and therefore switch-off times may be also up to 30% above the average value. Since the amplitudes plotted in Fig. 3 depend on the rate of switch-off of current in the transistor, overshoot amplitudes may fall to 75% of those in Fig. 3 in the worst cases. In order to ensure that the circuit will operate in such cases, the minimum overshoot in the average case must be fixed at 130% of 3·2 volts, i.e. 4·2 volts. This allows the amplitude with 30% increase in switch-off time to be at the minimum value of 3·2 volts.

Reference to Fig. 3 shows that if 4 mA is chosen as the emitter current, 2 mH is the least inductance which will give the necessary amplitude of overshoot.

The emitter current is not increased above 4 mA, so as not to present too great a load to the circuits which may be used to drive the stage.

(2.2.4) **Collector Current.**

The emitter-current supply having been fixed, the collector-current supply for J_1 may be determined.

J_1 is allowed to bottom when it conducts, the degree of bottoming being determined by the collector current. The resulting carrier storage has, however, very little effect on the speed of transition from one state to the other, owing to the circuit arrangement used.

Carrier storage time, when a bottomed transistor is switched off, is reduced considerably by taking the emitter potential negative with respect to that of the base. This is illustrated by the experiment shown in Fig. 5. Fig. 5(*a*) is a circuit in which

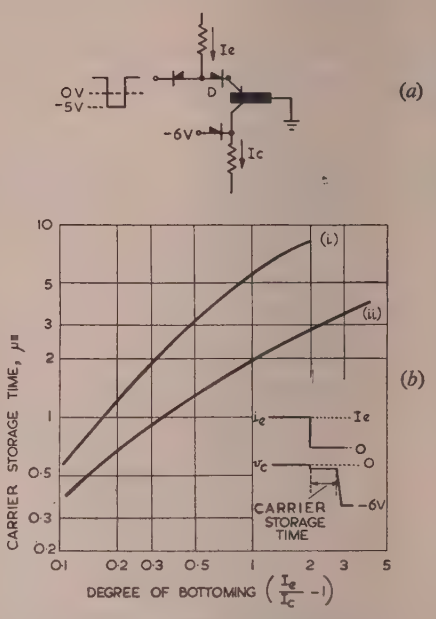


Fig. 5.—Carrier storage measurements.
(*a*) Circuit.
(*b*) Experimental results for type OC72.
(i) Diode D in circuit.
(ii) Diode D omitted.

emitter current is removed from a bottomed transistor after it has been applied for a long time.

Curve (i) of Fig. 5(*b*) represents carrier storage time, as measured with this circuit, plotted against the degree of bottoming, which may be represented as $(I_e/I_c - 1)$, I_e and I_c being the defined currents.

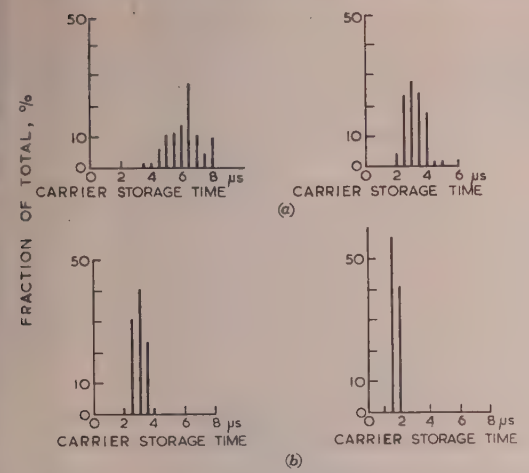
If the diode D [Fig. 5(*a*)] is short-circuited, the input pulse is allowed to pull the emitter potential negative. The emitter then acts as an extra collector, which results in an increase in the rate of withdrawal of carriers from the base region of the transistor. For a given degree of bottoming, therefore, the carrier storage time is reduced by omitting D, as shown by curve (ii) of Fig. 5(*b*).

This is similar to the way in which J_1 is connected in the binary-scaler circuit, and so a similar reduction in carrier storage time is to be expected during operation.

Fig. 6 shows the results of measurements on a batch of transistors, distributions of various carrier storage times being plotted for two values of $(I_e/I_c - 1)$, both with and without the diode D of Fig. 5(*a*).

The peaks may be seen to lie on, or very close to, the curves of Fig. 5(*b*), which are those for an average transistor. Although the spread is large when the diode is included, very little deviation from the peak occurs when the diode is omitted. Thus, although curve (i) of Fig. 5(*a*) may show a large variation between units, curve (ii) is much less variable. Increases in carrier storage time of 30% above the average may be encountered under the conditions of curve (ii), which must be allowed for when the circuit is being designed.

With a degree of bottoming $(I_e/I_c - 1)$ of 0·3, a carrier storage time of about 2 microsec in the worst possible case [corresponding to curve (i)] is indicated by Fig. 5(*b*). Allowing the emitter potential to move even a small amount negative results in a reduction in carrier storage time, so that operation tends towards the figures for curve (ii), and carrier storage time is about a microsecond even for a poor transistor.



g. 6.—Distribution of carrier storage times in a batch of transistors type OC72.

$I_e = 5$ mA.
Left-hand side: $I_c = 2$ mA.
Right-hand side: $I_c = 3$ mA.
(*a*) Diode D in circuit [Fig. 5(*a*)].
(*b*) Diode D omitted [Fig. 5(*a*)].

When an input pulse is applied to initiate a transition, it must st long enough to allow J_1 to become unsaturated before the ıange-over can take place. The minimum width of the input ulse is therefore set by the carrier storage time in J_1 at about microsec.

The presence of an inductance in the base of J_1 (the transrmer secondary winding) allows the input pulse applied to the nitters to be narrower than the minimum width defined above. he input pulse takes the emitter potentials negative and draws reverse emitter current from J_1, whose emitter now acts as a llector. This emitter current flows through the base inductance, nd when the input is removed, the overshoot of the base ductance takes the base potential positive with respect to the nitter, which is equivalent to lengthening the input pulse. hus, so long as hole storage is eliminated before the overshoot f the base inductance is completed, the input pulse may be made arrower than would be required if the base inductance were ot present. In fact, pulses of 0·5 microsec width are adequate r operating the circuit.

The degree of bottoming ($I_e/I_c - 1$) given above corresponds $I_c = 3$ mA, which is the value adopted in the circuit. The raph of Fig. 7 shows how switch-on time varies with the emitter

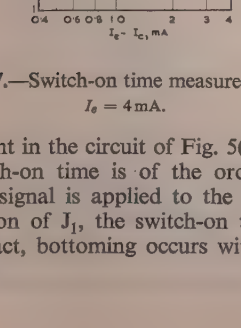


Fig. 7.—Switch-on time measurements.
$I_e = 4$ mA.

nd collector current in the circuit of Fig. 5(*a*). With the above urrents, the switch-on time is of the order of 2·5 microsec. Iowever, when a signal is applied to the base, as is the case uring the switch-on of J_1, the switch-on time is considerably educed, and, in fact, bottoming occurs within 0·5 microsec of pplying this pulse.

The switch-on pulse from the transformer is delayed upon the leading edge of the trigger pulse by the switch-off time of current in J_2, which is of the order of 1 microsec. J_1 therefore bottoms about 1·5 microsec after the application of the input pulse.

This time, and the time taken for the other transition, are the fundamental limits of resolving time in this circuit. Adjustment of circuit parameters to improve one of the above times is, in general, accompanied by a deterioration of the other, and no overall improvement is gained.

With a current of 1·5 mA flowing in R, and 3 mA current being required by the collector of J_1, a total of 4·5 mA is supplied to the junction of R and the collector lead.

(3) COMPLETE CIRCUIT, AND INTER-STAGE COUPLING

The complete circuit, with component values, is shown in Fig. 8(*a*). Typical waveforms are shown in Fig. 8(*b*).

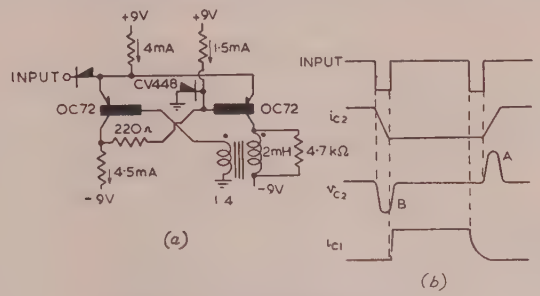


Fig. 8.—Complete binary scaler.
(*a*) Circuit.
(*b*) Waveforms.

Coupling from one binary scaler to the next is by means of a third winding on the transformer. By phase reversal or in-phase connection of this winding, the next binary may be actuated at either the switch-on (pulse A) or switch-off (pulse B) of the previous stage, giving flexibility in logical considerations. The ratio required for the tertiary winding is governed by much the same considerations as for the secondary winding: the loading due to the next stage must not be excessive, current must be stepped up into the next stage, and the voltage swing must be sufficient to produce current cut-off in the next stage. The ratio arrived at is the same as for the secondary winding.

The characteristics of the diode used for coupling to the emitters play an important part in the operation of the circuit. As has been shown, the input pulse must be shorter than the transformer pulse in any stage. Ordinarily this means that each transformer must have a greater inductance than that of the preceding stage. The resolving time of a chain of binary stages is not affected, since this is governed by that of the first stage, which has the smallest inductance. However, by choosing a diode with characteristics as shown in Fig. 9(*a*), it is possible to cascade identical stages, since the threshold characteristic of the diode in the forward direction effectively shortens the incoming pulse when J_2 is conducting, and, by causing a slight delay on its leading edge, causes the 'memory' pulse to be delayed, thus making it possible for the circuit to operate [Fig. 9(*b*)].

The threshold voltage is also essential in order to allow the emitter potential to rise slightly positive when J_1 is conducting, without having the diode (which is returned to earth) by-passing any of the supply current. This is possible so long as the voltage required to make the diode conduct is greater than the emitter-to-base voltage of a bottomed transistor.

The forward characteristics of silicon junction diodes make them especially suitable for use in this position, although germanium point-contact diodes are also satisfactory [Fig. 9(*c*)].

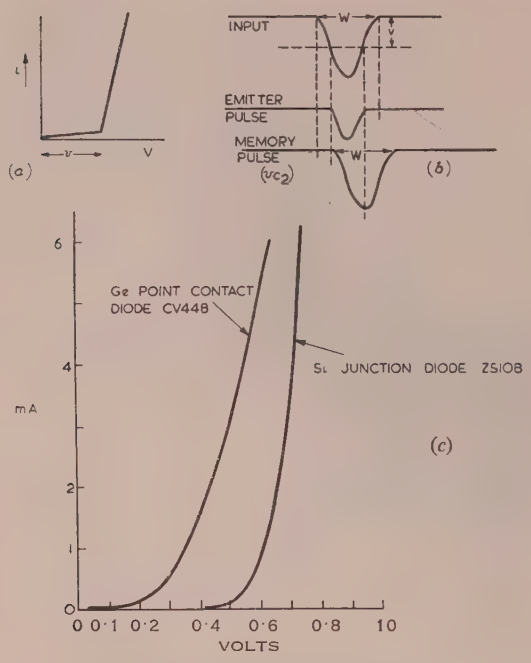


Fig. 9.—Effect of input diode.
(*a*) Idealized diode forward characteristics.
(*b*) Shortening of input pulse.
(*c*) Typical diode curves.

Representative germanium-diode samples show a current of 25 μA by-passed through the diode when J_1 is bottomed. This is too small to affect the operation of the circuit.

(4) INDICATION OF STATE

The current in the right-hand transistor may be used to provide an indication of state by inserting a resistance in series with the collector. Provided that this is small enough to prevent bottoming, its effect upon circuit operation is negligible [Fig. 10(*a*)].

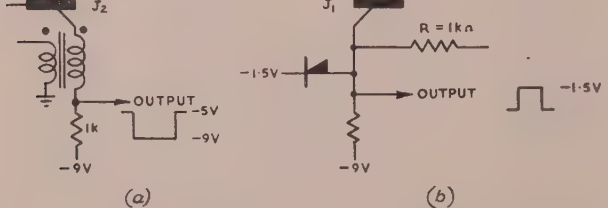


Fig. 10.—Methods of indicating state.

An indication in the opposite phase may be obtained from the collector of J_1 by a slight modification [Fig. 10(*b*)].

The resistance R must be increased to give a larger voltage swing at the collector, and the transistor must be prevented from bottoming, in order to give a clean falling edge to the collector waveform by eliminating hole storage. This is not important in the original circuit, since the collector current of J_1 does not have to fall to zero for the base potential of J_2 to be taken negative, and so carrier storage does not affect the operation. However, in order to provide waveforms to operate gates, etc., a clean falling edge is essential, and so bottoming is prevented in this case.

(5) DECADE SCALER

The basic circuit described above is used to form a decade scaler. The system to be used consists of a binary scaler and a scale-of-5 circuit in cascade, the binary scaler coming first in the chain. The resolving time of the decade is therefore that of the basic binary scaler (Fig. 11).

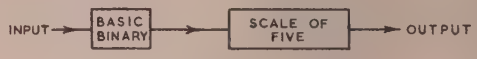


Fig. 11.—Block schematic of decade scaler.

There are many methods of connecting binary scalers to form a scale-of-5 circuit. One method* consists of a scale-of-4 circuit and an inhibiting circuit which prevents every fifth pulse from being counted.

Using the basic binary scaler described, a scale-of-5 circuit constructed on this principle consists of two unmodified binary scalers, B and C, and a third modified stage, A (Fig. 12).

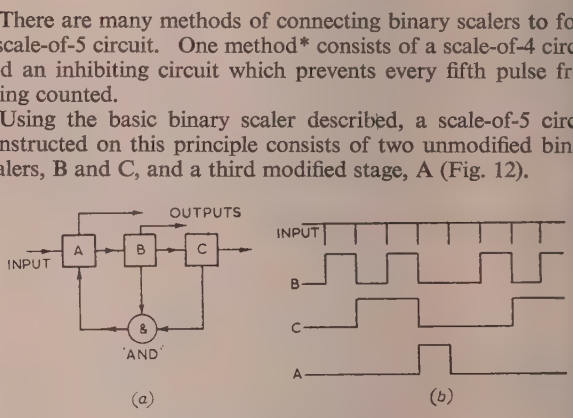


Fig. 12.—Scale-of-5 circuit.
(*a*) Block schematic.
(*b*) Waveforms.

Stage A is arranged so that input pulses are passed directly to the following stages B and C, which act as a scale-of-4 circuit. When an output is obtained from C, at the fourth input pulse this is fed back to stage A, which is then set so that the next (fifth) input pulse is not passed on to the scale-of-4 circuit. At the same time, this fifth pulse is used to restore the input stage A to its original condition, when the circuit is ready to repeat the process for the next five pulses. Thus a scale-of-5 circuit is obtained [Fig. 12(*b*)].

(5.1) Input Stage A

The input stage A consists of a binary stage with the a.c. feedback loop left open (Fig. 13). If the circuit initially has J_1 con-

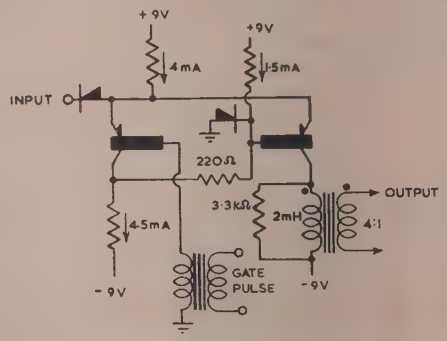


Fig. 13.—Input stage for a scale-of-5 circuit.

ducting, the first input pulse transfers the stage to its other stat with J_2 conducting. (This is the preferred state.)

Each input pulse then cuts off J_2, giving a negative-going puls in the transformer secondary winding, but does not cause

* Suggested by E. H. Cooke-Yarborough.

hange of state. Thus each pulse is passed on to the following ages. If now a negative pulse is applied to the base of J_1, the on-preferred state is set up. The next input pulse at the emitter ill transfer the circuit to the preferred state, but the output ulse which is obtained is of positive polarity, and does not ctuate the following stages. Thus the scaler is made to miss count.

(5.2) Gating

A simple method of deriving a gating pulse to change the ate of A after the fourth input pulse is to apply the output ulse from stage C, which occurs at the correct time (Fig. 14).

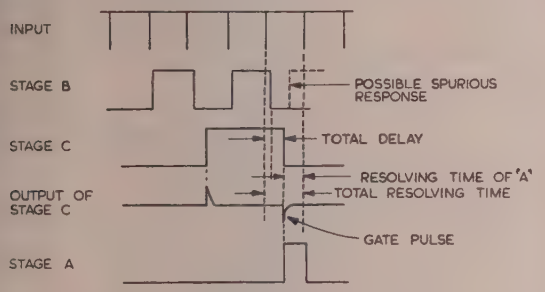


Fig. 14.—Simple gating waveforms.

nly when the current in J_2 (stage C) falls does a negative pulse ppear which may be applied to stage A.

Owing to a slight delay in each stage, however, the output of C will be delayed, and the change of state of A will also be elayed. Since A cannot be reset until a period equal to its time f resolution has elapsed, the resolving time of the complete cale-of-5 circuit is made up of the total delay time and the basic esolution time.

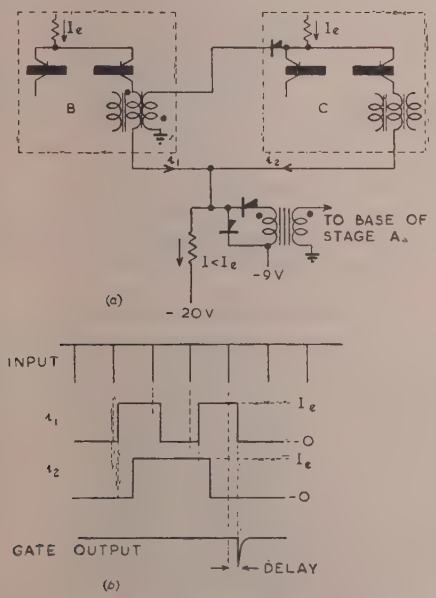


Fig. 15.—Improved gating.
(*a*) Circuit.
(*b*) Waveforms.

The total delay, which may be of the order of 2 microsec, makes this form of gating unreliable. The output pulse from A, which occurs at the fourth input pulse, will have decayed before the gate pulse is applied. When the latter does occur, stage A changes state, and in doing so produces a negative-going pulse in the output winding, which will be accepted by the following stage as another input pulse, and a spurious count will take place.

The delay on the gate pulse may be halved, to give an improved resolving time, and also to eliminate the possibility of a spurious count, by rearrangement of the logic.

The phase of coupling between B and C is reversed, so that the waveform of C is advanced by one period of input (Fig. 15). A coincidence gate combining the currents in B and C produces a gate pulse only when the currents both in B and C are zero. The output of the gate circuit is now delayed only by the delay time of one stage, which is of the order of 1 microsec. Spurious counting cannot therefore take place.

(6) OPERATIONAL TOLERANCES

The basic circuit which has been described is not unduly sensitive to changes in operating conditions. Changes in any one of the defined currents of $\pm 0{\cdot}5$ mA do not affect the operation, whilst simultaneous changes in the supply potentials of $\pm 20\%$ leave the circuit unaffected.

A wide range of input pulses may be accepted, ranging in width from 0·3 microsec to over 1 microsec at a minimum amplitude of 1·5 volts, the resolution being 2 microsec in each case. The scaler may be driven directly from the output pulses of a high-speed point-contact scaler[4] without modification.

(7) CONCLUSIONS

The results which have been described have been achieved with circuits which dispense with capacitors. It is felt that the use of capacitors in high-speed junction-transistor circuits has hitherto placed a limit upon the maximum speed of operation.

The use of inductances seems to present a useful technique in the design of transistor circuits in general.

It is of interest that the prevention of bottoming is not necessary in order to achieve the speed of which this circuit is capable. There is only very little decrease in minimum resolving time if bottoming is eliminated (about 0·3 microsec gain), and it is felt that this improvement does not justify the extra complication involved.

(8) ACKNOWLEDGMENTS

The authors would like to express their gratitude to Mr. E. H. Cooke-Yarborough for many helpful discussions.

(9) REFERENCES

(1) Linvill, L. G.: 'Nonsaturating Pulse Circuits using Two Junction Transistors', *Proceedings of the Institute of Radio Engineers*, 1955, **43**, p. 826.

(2) Wolfendale, E.: 'Two Bistable Circuits using Junction Transistors' (Mullard Research Laboratories Report No. T201R).

(3) Moll, J. L.: 'Large Signal Transient Response of Junction Transistors', *Proceedings of the Institute of Radio Engineers*, 1954, **42**, p. 1773.

(4) Chaplin, G. B. B.: 'A Point-Contact Transistor Scaling Circuit with 0·4 Microsec Resolution' (see page 505).

[The discussion on the above paper will be found overleaf.]

DISCUSSION ON THE ABOVE THREE PAPERS BEFORE THE JOINT MEETING OF THE MEASUREMENT AND CONTROL SECTION AND THE RADIO AND TELECOMMUNICATION SECTION, 27TH MARCH, 1956

Dr. A. R. Boothroyd: With reference to the paper by Dr. Chaplin, the high speed of operation is essentially due to the non-saturating nature of the 'on' states. Non-saturating bistable circuits are not new, and it is worth while to see in what way the present contributions represent an advance over previous circuits. In 1953 Carlson* described a two-transistor circuit utilizing the same basic d.c. arrangement as used in the paper; later he reported circuits operating at frequencies of several megacycles per second. In Reference 7 of the paper are described shift registers very similar to the present circuits, in which a common-emitter supply current defines the non-saturating 'on' state of one of several transistors. The circuits described in the paper are a considerable advance over these previous developments, however, owing to the method of triggering employed in conjunction with pulse-transformer intercoupling. The avoidance of coupling capacitors results in much greater speed of operation together with greatly improved reliability.

In non-saturating circuits the speed of operation is closely related to f_α, the alpha cut-off frequency of the transistors. The dependence of operation on f_α is not stated in the paper, nor is the minimum tolerated value of f_α stated. For example, the type 3X100N transistor can (according to the manufacturer's specification) have a value of f_α as low as 500 kc/s, which would certainly not allow operation of a scaling circuit at 2·5 Mc/s. What is the minimum value of f_α to be tolerated in the circuits described?

A closely related consideration, also dependent on f_α, is dynamic stability in the 'on' state. Much more is known about stability conditions than is indicated in the paper. Farley† has shown that, for stability,

$$C_e \leqslant \frac{1}{2\pi f_\alpha R_N}$$

where R_N is the magnitude of the emitter negative input resistance at the operating point and C_e is the emitter capacitance to earth. Thus f_α can be *too high*, causing instability. Also R_N can, if unduly large, cause instability; this emphasizes the importance of anomalous local regions with very high values of R_N which are present with many transistors, particularly ones with high values of f_α. Display of the emitter voltage/current characteristic of the transistors is recommended as a means of eliminating transistors suffering from such anomalies.

Mr. T. H. Walker: In the papers by Dr. Chaplin and Mr. Owens, the most interesting point is the attainment of high speed with bistable circuits. This is a distinct advance, since the fastest previous circuits have been based on monostable arrangements, which are by no means so flexible.

Have the authors had any opportunity of trying these circuits with transistors having higher cut-off frequencies? In the cases quoted it is demonstrated that the operating speed is limited by the current-gain cut-off frequency $f_{co\alpha}$, and it would be interesting to know whether, using the same circuit techniques, this still applies to more recent high-frequency transistor types.

In the junction-transistor circuit, one of the stable states depends on the emitter forward impedance of the junction transistor being less than the forward impedance of a diode. It depends on the forward drop of the emitter junction being less than the forward drop of the diode, and I would question whether this is entirely reliable. At the same time 100 mV in been taken as the design figure for the saturation-voltage dro of a junction transistor, and I would like to know whether th is considered to be a safe upper limit.

I should be very interested to learn whether the authors favo junction or point-contact transistors for counting circuits. this country, the speed factor is still in favour of the poin contact device, but that advantage may well disappear wi newer junction types, such as are appearing in the United State Do the authors feel, in spite of this, that the other advantag of point-contact transistors are really significant, and is the reliability adequate?

Have Dr. Franklin and Mr. James considered the possibili of using semi-conductor photosensitive devices. If this we practicable there might be advantages in terms of low-volta supplies and also absence of microphony.

Mr. E. Wolfendale: With reference to Fig. 5 of the paper Dr. Franklin and Mr. James, I have found that certain transisto with higher values of I_{co} tend to free run at higher temperature One way of improving the stability is to connect the seconda winding of T1 between the base and earth, thus eliminating t necessity for L1, and place the diode D1 in series with the emitte A resistance connected from the emitter to the negative supp line will bias D1 slightly positive and apply a small cut-off bi to the emitter of the transistor.

In connection with the papers by Dr. Chaplin, and Dr. Chapl and Mr. Owens, when scaling circuits are used in decade counte the limit to the resolving time is usually imposed by the puls shaping circuits required at the input to the first decade. Wh operating speeds have been achieved with pulse-shaping circui using the same transistors as those used in the decade?

With reference to the paper by Dr. Chaplin and Mr. Owe switching times of junction transistors are dependent on both and the base resistance. The transistor type OC72 used in t circuit has a much lower base resistance than is normal for low-frequency transistor. What speeds have been achiev using transistors with the same value of f_α but with the high base resistance.

I would question the statement by Dr. Chaplin that the poi contact transistor has a higher frequency response than t junction transistor. The development of the junction transist has produced high-frequency junction transistors with a high frequency response than the point-contact transistor.

Mr. K. Cattermole: In bistable circuits with non-regenerati switch-off, transition rates have been attained as high as tho usual in regenerative monostable circuits. Hitherto the latt have been not only faster but also more consistent, and I shou like to ask whether any selection of transistors is required. T transition times are not stated, but from Fig. 2 of the pap by Dr. Chaplin, they must be less than 0·1 microsec. D Chaplin mentioned that a point-contact transistor in his demo stration model produced 0·1 microsec pulses, and it would interesting to know the fastest transition observed.

The changes in emitter-base potential difference required operate a transistor are much smaller than the grid base of thermionic valve; even so, I prefer to use margins of about o volt to switch a transistor 'on' or hold it 'off'. Dr. Chaplin a Mr. Owens have reduced these to one- or two-tenths of a vo In the junction-transistor circuit, stability of one state depen on the forward voltage drop of a diode exceeding that of emitter, while the point-contact circuit requires tolerances on t triggering level and I_{co}, which I do not think could be maintain

* CARLSON, A. W.: 'High Speed Transistor Flip-Flops', AFCRC Technical Reports 53–16 and 53–16A, June, 1953, and March, 1954.

† FARLEY, B. G.: 'Dynamics of Transistor Negative Resistance Circuits', *Proceedings of the Institute of Radio Engineers*, 1952, **40**, p. 1497.

vith moderate rise of temperature, say to 35° C. Moreover, oltage margins which vary directly with supply potential (as in he classical junction-transistor bistable circuit with resistance ouplings from each collector to the opposite base) allow safe peration despite large supply fluctuations.

A thermionic valve is a high-impedance device with substantial elf-capacitance, but a transistor exhibits, and conveniently works nto, impedances of the order of the impedance of free space nd of practical conducting paths. It seems likely that eventually ransistors will become the first choice for high-speed digital peration, and Dr. Chaplin and Mr. Owens have made a signi- cant advance by devising circuits whose speed is limited only y the inherent properties of the device and not by capacitance n the external circuit.

Dr. Denis Taylor: The investigation into scaling circuits using oint-contact and junction transistors, respectively, was carried ut in order to discover how fast these circuits could be made vith the transistors which were currently available. There was ot necessarily any application in view.

With nuclear instruments we do not necessarily need the ighest speed. We have been carrying out investigations over a ong period into the required specification of our standard scaler, nd many discussions have taken place with the Americans and anadians, and also with members of the European Atomic nergy Society. Although there is no general agreement at resent, it seems that most people are thinking in terms of a tandard scaler with a resolving time of about 2 microsec, because his would meet the great majority of purposes for which scalers re wanted in nuclear work. There are, of course, cases where e need something faster, but this would be obtained by using pre-scaler in front of the standard scaler. So we have the hoice of using either the point-contact or the junction transistor.

I will leave aside the question of reliability—and this is, of ourse, coupled with the remarks of several speakers as to what ort of tolerances are reasonable and the spread of characteristics e can get with transistors of these types—and go on to the uestion of cost. A scale-of-ten instrument, using a point- ontact transistor circuit, and allowing for transistors, diodes, ransformers, and whatever else is required, costs about £16. he corresponding scaler using a junction-transistor circuit costs bout £14. If thermionic valves were used the cost would be 4 11s. So, at present, transistors cannot compete economically. t is difficult to make a true comparison, however, and I have ot, of course, allowed for the cost of the power supply. With hermionic valves a power supply of the order of 50 watts might asily be required, whereas 120 mW are required for the point-contact-transistor scale-of-ten instrument. The power supply is therefore much simplified, but the cost does not come down in the ratio of the two powers.

Mr. E. H. Cooke-Yarborough: Mr. Cattermole states that a reverse emitter bias of at least one volt should be applied to make sure of cutting off the emitter current. A very good feature of transistors, and particularly junction transistors, is that the characteristics are more calculable than in the case of valves. In particular, the emitter cut-off bias can be easily calculated. It can be shown* that the reverse bias voltage needed to reduce the emitter current of a junction transistor to zero is $\frac{kT}{e} \log_e (1 - \alpha)$. This means that if α is between 0·87 and 0·98, the emitter cut-off bias will lie between 50 and 100 mV. To this must be added the voltage drop produced by the flow of I_{co} in the base resistance, but this is usually unimportant, and so a reverse emitter bias of 200 mV gives an ample margin of safety.

Mr. J. N. Barry: In connection with the maximum counting rate of the scalers I would like to pose the question concerning the use of point-contact or junction transistors in a somewhat different manner from previous speakers. In the near future, junction transistors will, no doubt, be available having a cut-off frequency of the order of ten times that mentioned in the paper by Dr. Chaplin and Mr. Owens. If this could be assumed, would the authors still feel that the point-contact transistor has sufficient inherent advantages to be considered on its own merits for the types of circuit described?

I should like to associate myself with the comments made by Dr. Boothroyd regarding the anomalies which are seen in the characteristics of some point-contact transistors having high cut-off frequencies. It is worth mentioning another paper containing results of a detailed investigation into this subject,† and it is also worth adding that there is a correlation between such anomalies and the reverse-voltage transfer characteristic, which can be measured quite simply at a relatively low frequency.

I believe that the Americans have now officially adopted what may be called a more conventional symbol for the junction transistor than is used in the papers. It is similar to that used in the papers to denote the point-contact transistor, but it uses the double-headed arrow on the emitter lead, and has the complete symbol enclosed in an envelope. Although I agree that some differentiation is desirable between transistors which have inherent current gain (usually, but not necessarily, point-contact types) and those which do not, I do not think that the best solution is to perpetuate what is a relatively novel symbol to represent the junction transistor.

THE AUTHORS' REPLIES TO THE ABOVE DISCUSSION

Dr. E. Franklin and **Mr. J. B. James** (*in reply*): With regard o Mr. Wolfendale's remarks, it is agreed that the point-contact ilicon diode, shown as an example in Fig. 5, may not provide dequate stability with all types of transistor. In such cases the est course would appear to be the use of the silicon junction iode, which exhibits forward characteristics which are very uitable for the determination of the triggering level of the circuit. he arrangement proposed by Mr. Wolfendale would require he use of a very low forward-resistance diode, in order that the ransistor should remain in the 'on' condition until the voltage cross C had fallen to a low value. This would mean that a ermanium diode would be necessary, and the bias voltage eveloped across it would fall rapidly with increasing temperature. f the bias current were controlled by a thermistor, in order to ompensate for this, the power consumed in the bias circuit at he higher temperatures would be several times greater than that aken by the whole of the present circuit.

The possibility of using semi-conductor photo-sensitive devices, raised by Mr. Walker, has been considered for measurements of high radiation intensities. However, it is at present impracticable with the low radiation intensities encountered in geological and most radio-isotope applications, owing to the high noise level and small photosensitive area of presently available semi-conductor devices. The latter feature makes it difficult to obtain efficient light collection from the large-scintillation crystals required at low radiation intensities.

Dr. G. B. B. Chaplin (*in reply*): I am grateful to Dr. Boothroyd for drawing attention to the Technical Report on high-speed transistor flip-flops. This has not been available to me, but I presume the operating frequencies of several megacycles per second, quoted by Dr. Boothroyd, were achieved with American point-contact transistors having a considerably higher cut-off

* Shockley, W., Sparks, M., and Teal, G. K.: *Physical Review*, 1951, **83**, p. 158.
† Thomas, D. E.: 'Stability Considerations in V.H.F. Point-Contact Transistors', *Proceedings of the Institute of Radio Engineers*, **42**, p. 1636.

frequency than those available in this country. The purpose of the paper was to show how such cut-off frequencies can be utilized in scaling circuits without the external circuit imposing any appreciable limitation.

When operating the binary scaler at the maximum input pulse rate of 2·5 Mc/s the output waveform approximates to a 1·25 Mc/s sine wave, which is therefore the highest frequency which the transistor must amplify. Since the circuit requires a minimum current gain of 1·7, a transistor which has a higher current gain can tolerate a lower value of f_α; thus, for example, a transistor having $\alpha = 4$ need have a value for f_α of only 500 kc/s. Tests on the batch of 30 transistors showed a spread of f_α from 1·2 to 5·5 Mc/s, and all had a current gain greater than the minimum requirement of 1·7 at 1·25 Mc/s.

The dependence of stability on C_e and f_α, pointed out by Dr. Boothroyd, is, in fact, explained in Section 2.3, where it is shown that an adequate margin of stability was obtained with all samples tested.

Although the anomalous characteristics, mentioned also by Mr. Barry, do occur, they are not harmful unless they appear at the operating point. This is comparatively rare, but their elimination is taken for granted in all circuits, especially amplifiers, which do not specifically use these anomalies.

In reply to Mr. Wolfendale, it is not possible for an individual stage of the decade to be used as a pulse-shaping circuit if the subsequent decade is to be operated with its minimum resolving time, since the pulse shaper must operate at double the frequency of the decade. If the minimum resolving time of the decade is to be realized, the shaping must be done in valve circuits.

In reply to Mr. Cattermole, the switch-off time under non-regenerative conditions, assuming an instantaneous input step, is fundamentally the shortest that can be achieved, and the circuit approximates to these conditions, since, by definition, the input pulses are shorter than the resolving time. The 0·1 microsec pulses used for driving the demonstration scaler were derived from a selected transistor having a cut-off frequency of 5 Mc/s.

The circuit of Fig. 5 does not require close tolerances on either i_{co} or the triggering level, since it uses the defined base-current and catching-diode technique. The base-current supply, I_b, should be chosen to be greater than the highest expected value of i_{c0}, and the input trigger pulse should merely be adequately larger than the forward conduction voltage of two point-contact diodes in series, 2 volts being sufficient, as explained in the paper. It follows that the triggering level is also substantially independent of supply voltage.

I agree with Mr. Wolfendale's contention that the new experimental junction transistors have higher frequency responses than the point-contact type of several years' standing, but unfortunately they are not yet generally available.

Messrs. Barry and Walker both raise the interesting question whether scaling circuits using these faster junction transistors will be preferable to those using point-contact types. The experimental high-frequency British junction transistors allow simple binary scalers to resolve about 0·2 microsec, while resolving time shorter than 0·1 microsec can easily be achieved with the Ameri can surface-barrier transistors. If the cost of a British high frequency junction transistor is similar to that of a point-contac type, the junction circuit is preferred unless a large output voltag swing is required.

Dr. G. B. B. Chaplin and **Mr. A. R. Owens** (*in reply*): We do not think that the question of base resistance, raised by Mr Wolfendale, is significant in the case of the junction-transisto binary circuit. With a transistor which is being switched by means of a step of voltage applied between the base and emitter the switching speed will depend upon the base resistance of th transistor. However, with the binary scaler described, switchin is achieved by means of currents gated into the emitters, and so the operation is quite independent of base resistance. Substitu tion of transistors with base resistance four times that of th type OC72, but otherwise very similar, increased the resolvin time by less than 5%, which seems to bear out the above reasoning.

Messrs. Cattermole and Walker have entertained some doubt as to the reliability of the voltage margins which we have allowed for maintaining the stable states. However, as Mr. Cooke Yarborough points out, these margins may be readily and accurately calculated in the case of junction transistors. The bia required to cut off J_1 (Fig. 1) is always less than 100 mV at ordinary operating temperatures, and the figure of 200 mV which Mr Cooke-Yarborough quotes provides an adequate margin o safety. The reverse bias which is being applied to J_1 is th difference between the forward drop of D_1 and the conductin emitter-base potential of J_2, and actually increases with a rise i temperature. The forward potential drop at the emitter of J is calculated to be never greater than 200 mV at room temperatur for type OC72 transistors in this circuit, and this figure decrease rapidly with increasing temperature. The voltage drop acros D_1 also decreases with temperature, but at a lower rate, s that the difference increases with temperature. Choosing diode type with a forward drop greater than 0·4 volt at roon temperature ensures that the 200 mV reverse bias which i required is always obtained. The figure of 100 mV which wa quoted for the collector-to-base saturation voltage of J_1 is also maximum figure, and both observation and calculation indicat that, in practice, the actual value will be less than this.

Mr. Barry raises the question of symbols to be used to denot junction transistors. We find that the symbol used in the pape is extremely useful, in that it combines a readily recognizabl and easily drawn form with a useful physical picture of th transistor. It is worth recalling that the point-contact transisto symbol also represents a physical form. Although, as Mr Barry points out, the Americans have a standard symbol for th junction transistor, the need for a more definite differentiatio between the two symbols seems to have been recently appreciated to judge from the diversity of junction-transistor symbols appea ing in current American literature.

621.396.969

Paper No. 1940 R
Nov. 1955

FREQUENCY-MODULATION RADAR FOR USE IN THE MERCANTILE MARINE

By D. N. KEEP, B.Sc.(Eng.), Associate Member.

The paper was first received 29th April, and in revised form 16th August, 1955. It was published in November, 1955, and was read before the RADIO AND TELECOMMUNICATION SECTION *7th March, and the* NORTH-WESTERN RADIO GROUP *14th March, 1956.)*

SUMMARY

The principles of f.m. radar are outlined and a comparison is made between pulse and f.m. techniques, particularly with respect to the requirements of the merchant service. It is concluded that multi-gate f.m. radars are too complex for this application and methods are outlined for overcoming the inherently low scanning rate of single sweeping-gate systems. Equipment is described which has an aerial beamwidth of 1·7° and a rotation rate of 10 r.p.m. with a fractional range resolution of 1/30.

The future of f.m. radar for mercantile marine use is critically examined, the conclusion being that it will be most useful where very-short-range high-resolution pictures are required. Before such equipment is economically available further developments in transmitting valves must take place.

LIST OF PRINCIPAL SYMBOLS

f_0 = Frequency deviation of the transmitter.
R = Range of target.
c = Velocity of electromagnetic waves.
df/dt = Slope of frequency modulation.
δR = Range resolution, absolute.
$1/n$ = Fractional range resolution = $\delta R/R$.
f_m = Modulation repetition rate.
f_e = Echo beat-note.
f_d = Doppler frequency shift.

(1) INTRODUCTION

(1.1) Definition

Frequency-modulated radar may be defined as radar in which continuous-wave transmission is frequency-modulated in a known manner in order to obtain range information.

(1.2) Historical

The earliest use of this type of system was by Appleton and Barnett in 1924[1] when they wished to obtain evidence of the existence of the ionosphere.

Since then, the principles of f.m. radar have been restated several times for use in aircraft altimeters, and this type of altimeter came into widespread use during the Second World War.[2,3] Little work was done, however, on f.m.-radar multi-target systems until about 1940, when some development was undertaken in this country[4,5] and in the United States.[2]

As a result of this work the Ministry of Transport and Civil Aviation became interested in the system for mercantile marine applications, and in 1948 the Royal Naval Scientific Service undertook some development on behalf of the Ministry. Examination of a single sweeping-gate radar with orthodox design parameters showed that such a system has a data rate which is far too low for navigational purposes.

It became evident, however, that it could be designed to give reasonably high data rate with a proportional sacrifice of sensitivity. The paper describes the design and performance of such an equipment.

Mr. Keep is in the Royal Naval Scientific Service.

(2) OUTLINE OF THE SYSTEM

In f.m. radar, c.w. power is transmitted which is frequency-modulated, a sawtooth waveform being normally used [Fig. 1(*a*)].

The energy reflected from a target differs in frequency from

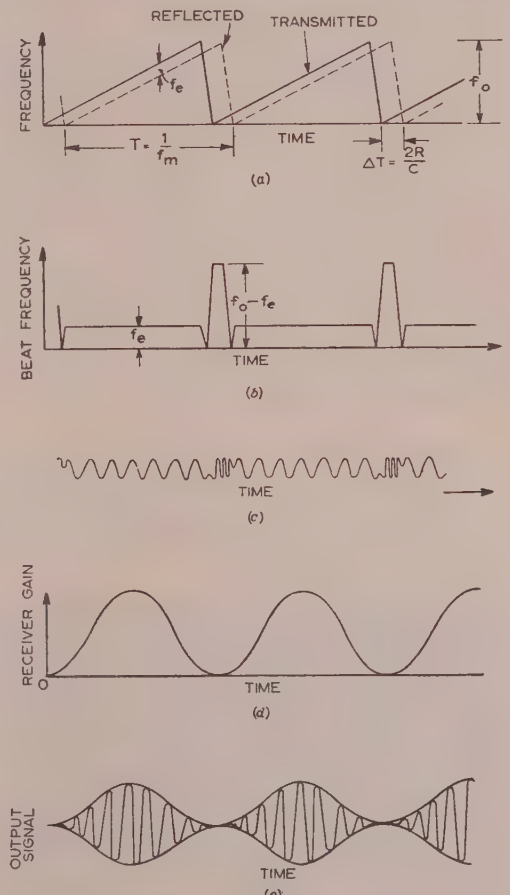


Fig. 1.—Method of obtaining echo signals in f.m. radar.
(*a*) Frequency variation of signals.
(*b*) Beat frequency for one target.
(*c*) Beat-frequency waveform, exaggerated.
(*d*) Receiver gain modulation.
(*e*) Beat frequency after gain modulation.

that transmitted, by an amount $2(R/c)df/dt$, R being the range of the target, c the velocity of electromagnetic waves and df/dt the slope of the sawtooth modulation. This difference-frequency is extracted by beating the received energy with a fraction of the transmitted power.

Owing to the repetitive modulation, the spectrum of possible beat notes from stationary targets consists of a set of single frequencies separated by a frequency f_m, the modulation repetition

rate, the energy from most targets containing two or three significant components.[6]

In order to distinguish between targets at different ranges, a set of frequency filters of bandwidth f_m c/s, with centre frequencies separated by the same amount is required. Such a set of filters is known as a multi-gate receiver.

Theoretically, the range resolution is given by $c/2f_0$ (where f_0 is the frequency deviation of the transmitter) as in the pulse-radar case when similar simplifications are applied. However, in the f.m. radar derivation it is assumed that the linearity of the frequency-modulation characteristic is adequate, and this may be a severe practical limitation. An assessment of the effects of non-linearity may be obtained by the method outlined in Reference 6.

On the flyback of the frequency modulation, false or unwanted signals are likely to be produced, particularly as the tendency of the high value of df/dt will be to bring large, short-range objects into the range of beat frequencies in use. To avoid this the receiver gain is suppressed during the flyback[4,7] [Fig. 1(*d*)].

To complete this Section mention must be made of the Doppler shift frequency f_d. For a relative velocity of some 30 knots the Doppler shift at X-band will be about 900 c/s. This shift frequency directly adds to or subtracts from the beat frequency (depending on the relative motion of the target and the sign of the slope of the frequency sweep) and may thus render the range information inaccurate.

Of the various methods of separating the two frequency-shift components, the simplest is probably the use of symmetrical triangular modulation.[2] However, it will become evident that the Doppler shift is negligible in the designs of radar considered in the paper, and extraction of velocity information is not normally required in the mercantile marine.

(3) COMPARISON BETWEEN PULSE AND F.M. RADARS

It can be shown theoretically that f.m. radar with a multi-gate receiver (i.e. one receiver filter for each element of range) is equivalent in performance to pulse radar having the same mean power and utilizing the same bandwidth, the aerial beamwidths, scanning speeds, etc., being equal.[6]

From the above outline a number of important differences are apparent between the pulse and f.m. radar rechniques. These are:

(*a*) The f.m. radar is transmitting c.w. power and is not therefore concerned with high peak powers. The importance of this becomes overwhelming when the equivalent pulse set has peak powers approaching, or greater than that which the waveguide system can carry without breakdown.

(*b*) It is no longer possible to use time duplexing of one aerial since c.w. transmission is used. It is possible to use separate transmitting and receiving aerials or circular polarization duplexing.[8] If circular polarization is used for duplexing, rain clutter can no longer be rejected by the same means.

(*c*) Echoes from all ranges are present in the early stages of the receiver simultaneously, and if the total energy is large the receiver may become saturated, leading to cross-modulation and distortion. Rain clutter from the very shortest ranges may be severe, and use of separate transmitting and receiving aerials may be essential in order to decouple the near zones. Sea clutter is obviously not such a serious problem.

(*d*) In f.m. radar the bandwidth necessary to attain the range resolution specified is only required at the r.f. frequencies, whilst in pulse radar the full bandwidth is required in the i.f. amplifier. This sets a practical limit to the range resolution attainable with pulse techniques. However, it must not be forgotten that, to take advantage of the wider bandwidths with f.m. technique a high degree of modulation linearity is required (see Section 2).

(*e*) For optimum utilization of f.m. radar signals, the receiver must consist of a set of parallel filters, the required number being equal to the fractional range resolution. Furthermore, the receiver outputs must be combined in some manner if a display of the conventional type is required.

(*f*) The transmitting valve may act as the local oscillator provid that it is not excessively noisy, and this leads to some saving in th r.f. components of the equipment. However, alternative systems which a local oscillator is kept in step with the transmitting valv must not be overlooked.

(4) CONSIDERATION OF F.M. TECHNIQUES FOR MERCANTILE MARINE USE

In the mercantile marine, navigational radar systems us comparatively small peak powers. Nevertheless the use of f.m techniques should eliminate the need for moderately high powered modulators and, since the transmitter can also act a the local oscillator, only one X-band valve may be required.

The f.m. radar duplexing problem is not serious in this fiel since cheese aerials are reasonably satisfactory, and the increas in bulk in using one on top of the other is small.

The possibility of attaining high range-resolution is ver pertinent to the navigational field, particularly for short-rang working, such as navigating narrow waterways. The necessit for using a multi-gate receiver is, however, a serious drawbac since it will result in a large and complicated set compared t normal pulse-radar equipment, particularly when the fractiona resolution is high. This consideration led to the exploration an use of the single-sweeping-gate f.m. radar system, which sacrifice performance for simplicity of design.

(5) SINGLE-GATE F.M. RADAR SYSTEMS

In a single-gate system, the beat-notes corresponding to th ranges required are either scanned across a single fixed-frequenc gate or, alternatively, the filter frequency is varied. By thi means the reciever is considerably simplified and it is quite eas to form a normal p.p.i. picture. The cost, however, is heavy Since range is now being scanned it is obvious that th overall data rate must decrease or the sensitivity be lowered A system along these lines was tried, and using the values c modulation repetition rate and aerial beamwidth that are norma in pulse radar (500 c/s and 2°) it was found that with a rang resolution of 1/30 an aerial scanning rate of approximatel 1½ r.p.m. only was achievable, and it was therefore necessary t scan the aerial over a restricted sector. It was quite clear tha the performance was inadequate for navigation and anti-collisio use, and attention was given to increasing the aerial scannin rate without unduly complicating the equipment.

Other forms of single-gate f.m. radar have been considere from time to time.[2] In most of these, however, the absolut range resolution varies over the range scale, and this is considere to be an undesirable feature.

(6) METHODS OF IMPROVING THE SCANNING RATE O SINGLE-GATE F.M. RADAR SYSTEMS

The scanning rate, once beamwidth and fractional rang resolution have been determined, is proportional to the bandwidt of the receiver gate, which for optimum signal/noise ratio shoul be matched to the modulation repetition rate. To increase th data rate the modulation repetition rate can be increased and th i.f. bandwidth held at its optimum figure, or, alternatively, th bandwidth increased whilst the modulation repetition rate is hel constant.

Increase in the modulation repetition rate is limited by th occurrence of second-trace echoes and is therefore mainl applicable to the shortest range scales. The second-trace echo are caused by the returned energy from a distant target beatin with a subsequent cycle of the modulation.

The use of a receiver bandwidth which is considerably large than the optimum is also limited. The receiver must now explo several elements of range during one period of the frequenc

modulation of the transmitter, and none of these may be lost during the period that the receiver is desensitized for the flyback of the modulation. Furthermore, a proportionately wider frequency deviation of the transmitter is necessary in order to maintain a given range resolution.

(7) A CONTINUOUS-SCANNING SINGLE-GATE F.M. RADAR

By progress along these lines it was concluded that an aerial scanning rate of 10 r.p.m. could be attained with a range resolution of 1/30, and with reasonably good sensitivity.

The gate bandwidth of the set designed is 12 kc/s, and for range scales below 1 000 yd a modulation repetition rate of 10 kc/s is used. For ranges above this the modulation repetition rate is reduced to 2·5 kc/s, in order to avoid second-trace effects. Twin cheese aerials were used in the model, although their horizontal beamwidth of 1·7° is rather smaller than the optimum (about 3·5° is required to make the mean tangential resolution equal to the radial resolution). This permits only 4–5 echoes per point target, which necessitates some tangential modulation of the time and can be fed to a normal p.p.i. display, the time-base of which is locked to the panoramic sweep. It is necessary to reject the lowest part of the range scale since harmonics of the gain modulation and also of the residual amplitude modulation of the klystron are extremely troublesome here and constitute a kind of ground-wave. Within the range 50 kc/s–1 Mc/s the sensitivity is limited only by random noise from the r.f. crystal input.

The transmitter modulator unit, panoramic receiver and necessary power units are housed in a two-drawer unit measuring 24 × 20 × 20 in. The display unit consists of a 12 in diameter p.p.i. display.

The experimental set contains 75 receiving-type valves and one microwave transmitting valve; this number could be reduced in a production design, leading to an even more compact set.

(8) PERFORMANCE OF THE SET

Fig. 3 shows a typical long-range picture obtained with the set, using an experimental klystron valve with an output power

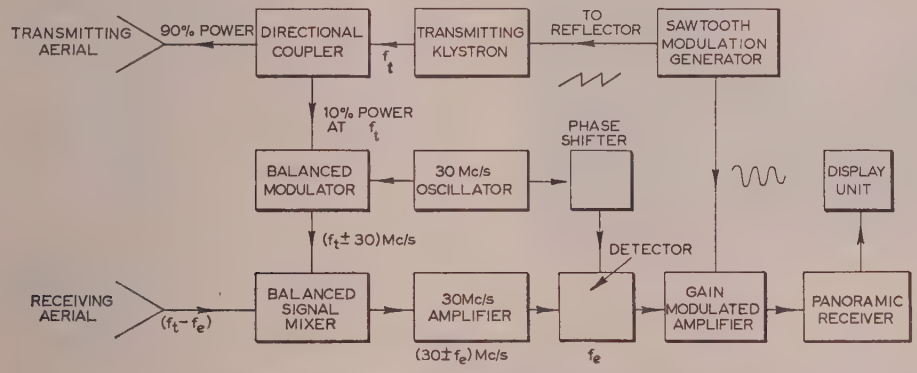


Fig. 2.—Simplified diagram of f.m. radar set.

cathode-ray-tube beam in order to avoid spoking on the display. A block diagram of the complete set is given in Fig. 2.

A 30 Mc/s sub-carrier is used in order to take advantage of the low noise factor of mixer crystals at the higher intermediate frequencies. A balanced modulator and a balanced signal mixer are necessary to eliminate amplitude modulation of the transmitter from the receiver input, and these, together with the 30 Mc/s oscillator, directional coupler and transmitter klystron, are housed with the 30 Mc/s head amplifier in the r.f. box beneath the aerials.

The beat-notes are obtained as modulation products of 30 Mc/s, and after some amplification the signals are detected in a linearized detector.

All signals are simultaneously present in the head amplifier and it is therefore essential that it is not forced into the non-linear region by very large signals. The signal level at the detector is consequently very low, the head-amplifier gain being just sufficient to make its input noise the determining level for the receiver noise factor.

After detection, the signals are gain-modulated in a balanced modulator and then passed to the panoramic receiver, where they are analysed sequentially. The panoramic receiver consists of a local oscillator electronically swept over the frequency range 3·05–4 Mc/s about 150 times per second (sawtooth waveform). The output is fed to a balanced mixer where it is mixed with the incoming signals which are in the frequency range 50 kc/s–1 Mc/s. The output is then presented to a 3 Mc/s i.f. amplifier having a bandwidth of 12 kc/s. In this way the signals are separated in

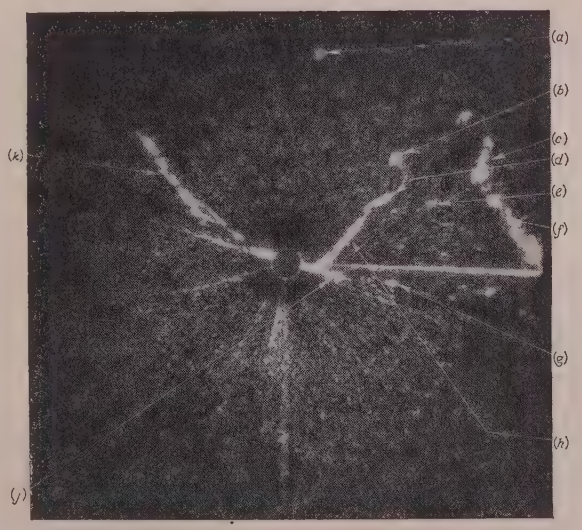


Fig. 3.—P.P.I. photograph, 9 nautical miles range scale, 80 mW transmitted, radar set sited at Southsea.

(*a*) Nab Tower.
(*b*) Ship.
(*c*) Isle of Wight hills.
(*d*) Horse Sand Fort.
(*e*) No Man's Fort.
(*f*) Hills.
(*g*) Fort.
(*h*) Defence boom.
(*j*) Pier.
(*k*) Coast-line of Hayling Island.

of approximately 80 mW. The range scale shown is 9 nautical miles, with a range resolution of 1/30, and modulation repetition rate 2·5 kc/s. The picture is one of the sea area off Southsea, Hampshire, the radar set being situated close to the beach. The inland area (lower part of picture) has no echoes showing, since the radar was screened in this direction by buildings, the echoes from which are rejected by the receiver since they are at a very short range.

Fig. 4 shows a very short-range picture obtained with a K302

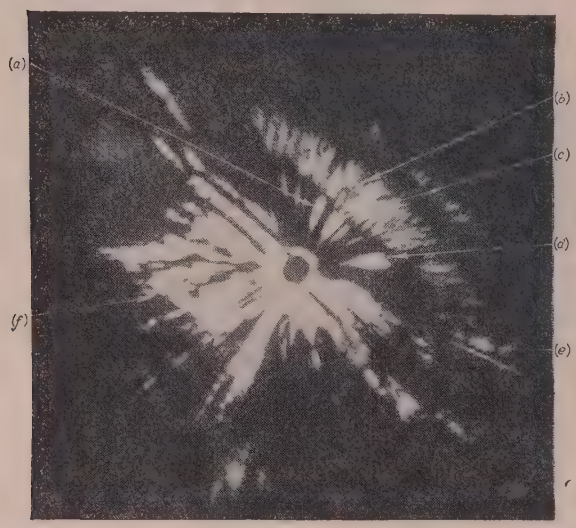


Fig. 4.—P.P.I. photograph—100 yd range scale, 15 mW transmitted.

(*a*) Lamp-post.
(*b*) Shingle bank of beach.
(*c*) Seat.
(*d*) Man walking along promenade.
(*e*) Lamp-post, 70 yd range.
(*f*) Buildings.

klystron (output power approximately 15 mW, modulation repetition rate 10 kc/s) with its f.m. characteristic linearized by feeding it into a frequency-sensitive load of the correct magnitude. Under these conditions a frequency sweep of 36 Mc/s was obtained within limits of linearity of ±5%. This linearization system was first used successfully in one of the earliest f.m. systems built in this country.[12] The range scale is approximately 100 yd and the radar set is sited as for the longer range-scale picture. However, in this test practically the entire range scale was over land and it was found necessary to reduce the amplifier gain considerably in order to pick out any targets in the land clutter.

The fractional resolution on this range dropped to about 1/10 mainly because the klystron sweep is not sufficiently wide or linear to obtain the full 1/30 range resolution of the panoramic receiver. Nevertheless, this range scale was used for demonstration purposes because it brings into view a number of discrete targets that are readily recognizable, and human targets in motion could be easily picked out.

An important point about the short-range performance is the very high brilliance obtained on the display, since the cathode-ray-tube beam is sweeping continuously at a fairly low speed. The minimum range obtainable was about 10 yd. To some extent this performance is aided by the vertical beamwidth of the aerial (±12·5°) which reduces signal levels at the shortest ranges.

Sensitivity trials were carried out and the performance of the set was compared with that of a typical pulse radar. The performance of the f.m. set was found to be 30 dB down, 18 dB of this being due to lower mean power and the remaining 12 dB to the large receiver bandwidth of the f.m. radar, which was non-optimum in comparison with the pulse set.

Very little wave-clutter was observed during extended trials of the equipment. This is readily understandable since the equipment was about 30 dB down in performance compared with a typical pulse set, and was, furthermore, situated about 100 yd from the edge of the sea at high tide. Examination of wave-clutter results obtained with pulse sets indicate that the clutter amplitude will not reach excessive values provided that the c.w. power transmitted does not appreciably exceed the values at present used in normal pulse navigational radar—i.e. about 1 watt. No rain clutter was observed on the set.

(9) CONCLUSIONS

It may be concluded that a single sweeping-gate f.m. radar system may be designed to have a moderate fractional range resolution and a reasonable aerial rotation rate.

On a range scale of 1 000 yd a resolution of 1/30 is equivalent to 33 yd, which can be achieved with a good pulse radar. On longer-range scales the f.m. radar resolution is worse than that attainable by pulse techniques. On shorter-range scales it may be better, but before this can be exploited fully a much wider linear frequency deviation must be obtainable from the transmitting valve.

The brilliance of the short-range pictures obtained by f.m. radar is high, and it appears that the technique could be exploited for applications requiring very short-range scales. For this purpose the equipment can be considerably simplified by eliminating the 30 Mc/s sub-carrier, the loss of 10 dB or so in sensitivity being unimportant at these short ranges. The slant range distortion may become large at very short ranges, depending upon the relative heights of the targets and aerials. Provided that these can be considered constant, the distortion can be eliminated by suitable adjustment of the time-base law.

Large areas of the close-range picture may be obstructed when the set is fitted on board a ship. If this is sufficiently serious two scanners will have to be used, one mounted on each side of the ship.

It is possible to envisage an f.m. radar set designed primarily for use at short ranges but having long-range scales and performance sufficiently good for ordinary navigation should the main pulse-radar set fail.

(10) FUTURE DEVELOPMENT OF F.M. RADAR FOR THE MERCANTILE MARINE

Future developments will be mainly concerned with the problem of obtaining suitable transmitting valves, and with improvement of the radar data rate, since the fractional range resolution achieved is rather low.

Linearization of the klystron tuning characteristic by external loading is not a practical possibility outside the laboratory, since the setting-up is somewhat critical and the actual tuning point very much so. However, linearization can be incorporated into the valve itself by coupling into the normal klystron resonator a synchronously tuned secondary resonator. Work along these lines has been carried out at S-band frequencies in the United States.[10] Continuous-wave valves now becoming available may well be suitable for this application. Carcinotron valves[11] have extremely good tuning characteristics and it should be only a matter of time before they are available for X-band frequencies. Alternatively, the VX3238 c.w. magnetron can be pulled in frequency over a wide range. Mechanical methods of pulling are out of the question at the high modulation frequencies required, but it may be possible to develop an electronic method utilizing ferrites.[9]

The more serious and fundamental problem is that of increasing the fractional range resolution without decreasing the aerial

anning rate. Some improvement can be expected by further djustment of the design parameters of the set, but such changes e inevitably limited by sensitivity and second-trace echo equirements and also by design complications.

The limitation is introduced, of course, by the single-sweeping-ate receiver, and the ultimate answer would be the use of a ulti-gate receiver. Until such a receiver becomes an economic ossibility, the hope of a moderately large improvement appears lie in the use of several parallel gates each sweeping a fraction f the range scale, thus going some way towards the goal.

The increase in size of the equipment due to the use of several arallel gates is fairly small, but it does introduce the problem of econstituting the several outputs to give a conventional p.p.i. icture, and the cost of this may be relatively high. If the number f parallel channels is small, the most attractive proposition, heoretically, is to use a multi-beam cathode-ray tube, since this ould preserve completely the low sweep-speed of the beam. A uitable four-beam tube is available in this country but not in zes larger than 6in diameter, and its cost is very high.

It would be possible to store the information from each channel nd read it back in the correct sequence. The speed of operation ould be high, however, so that this solution is likely to be no ss expensive than the first.

An alternative method is to use a normal p.p.i. display in onjunction with an electronic switch. The writing speeds would e higher than in the simple single-channel system, but with a mall number of channels they would not be excessive. This olution is likely to be the most economic but may lead to a onsiderable increase in the size of the set.

(11) ACKNOWLEDGMENTS

The author is indebted to the Admiralty and to the Ministry of ransport and Civil Aviation for permission to publish the paper. he conclusions drawn are those of the author and do not necessarily reflect the views of the Departments. The author wishes to acknowledge the invaluable contributions of Mr. J. Croney and Mr. S. de Walden, and also to thank those of his colleagues who were associated with the construction and testing of the equipment described.

(12) REFERENCES

(1) Appleton, E. V., and Barnett, M. A. F.: "On Some Direct Evidence for Downward Atmospheric Reflection of Electric Rays," *Proceedings of the Royal Society*, A, 1926, **109**, p. 554.
(2) Luck, D. G. C.: "F.M. Radar" (McGraw-Hill, 1949).
(3) Smith, R. A.: "Radio Aids to Navigation" (Cambridge University Press, 1947).
(4) Standard Telephones and Cables Ltd.: British Patent 581169.
(5) Rust, N. M., and Partington, G. E.: British Patent 621846.
(6) Gnanalingam, S.: "An Apparatus for the Detection of Weak Ionospheric Echoes," *Proceedings I.E.E.*, Paper No. 1670, July, 1954 (**101**, Part III, p. 243).
(7) Marconi's Wireless Telegraph Co. Ltd.: British Patent 647583.
(8) Ramsay, J. F.: "Circular Polarization for C.W. Radar," Proceedings of a Conference on Centrimetric Aerials for Marine Navigational Radar" (Ministry of Transport, 1952).
(9) Hogan, C. L.: "The Microwave Gyrator," *Bell System Technical Journal*, 1952, **31**, p. 25.
(10) Reed, E. D.: "A Coupled Resonator Reflex Klystron," *ibid.*, 1953, **32**, p. 716.
(11) Warnecke, R., and Guenard, P.: "Some Recent Work in France on New Types of Valves for the Highest Radio Frequencies," *Proceedings I.E.E.*, November, 1953 (**100**, Part III, p. 351).
(12) Private communication from Marconi Research Laboratories.

DISCUSSION BEFORE THE RADIO AND TELECOMMUNICATION SECTION, 7TH MARCH, 1956

Mr. N. M. Rust: I was particularly impressed by the 9-mile-ange p.p.i. picture (Fig. 3), which shows much what I would xpect for the conditions given; on the other hand, I was dis-ppointed with the short-range picture (Fig. 4), where the range as only 100yd. Allowing for the fact that the linearity btainable with the valve used for a sweep of 36Mc/s was not etter than $\pm 5\%$, and for other possible causes, Fig. 4 does not epresent the best that can be obtained with f.m. radar under the est conditions.

Although it is clear from the paper that the author appreciated heoretically that 'gain modulation' has a purpose to fulfil other han the suppression of spurious responses that would otherwise e produced by nearby echoes during the sharp downward stroke f the sawtooth swing, I feel that the practical importance of his other function for short-range working may not have been learly realized. Fig. 1(*c*) shows clearly how the beat note is hopped by the sawtooth swings, and it will be noticed that there s no coherence of phase between the portions for different nodulation cycles. This series of non-coherent beat-note spasms stablishes a frequency spectrum of harmonics of the modulation epetition frequency, and, if gain modulation is not employed in he manner illustrated in Fig. 1, this spectrum spreads out very onsiderably on either side of the mean beat-note frequency, in he same manner that a pulse spectrum spreads. To prevent this pectrum spreading unduly at the skirts it is most important to hape the amplitude of the beat-note spasms. The Fourier-ransform relationship between amplitude envelope and sideband pread is the same as for a pulse, and is analogous to the relation-ship between the amplitude distribution over an aerial aperture and the resultant aerial polar diagram.

Constant distribution—using the aerial analogy—across an aperture gives a sharp beam with big side-lobes. A distribution tapered symmetrically about the centre of the aperture increases the beam width but reduces the side lobes. In the same way, incorrect or insufficient gain modulation gives apparently sharper range resolution for distant objects, where the weaker remote side-lobe components will not display. This is very largely due to the fact that the light response is not linear in the cathode-ray tube. On the other hand, at short ranges these weaker wide-spreading frequency components may have a serious effect on the display definition, owing to the much bigger signal strength.

An examination of Fig. 4 would lead to the belief that this factor may partly account for the disappointing resolution obtained. Looking at the two lamp-post paints at (*a*) and (*e*), (*a*) is close in and (*e*) is near the end of the range at 70yd. If non-linearity were the principal contributory cause of bad resolution, one would expect that (*e*), the long-range one, would spread considerably more than (*a*), because in the system used the spread due to this cause is proportional to the range. It would therefore seem likely that spectrum spread, owing to a gain-modulation adjustment which was not best adapted to short-range working, might account, to some extent, for poor resolution.

Although I have drawn attention to the importance of correct gain-modulation adjustment, I do not want to infer that linearization is not important. The author refers to linearity control by the use of frequency-sensitive loading, usually carried out by

stub adjustments on a waveguide. Modern methods of linearization, using ferrite modulators and feedback correction, will make such experimental methods completely out of date, and we would not contemplate, in the future, a set in which it was necessary to make running adjustments.

With regard to future developments, there is one very important point which should be borne in mind. When using high-definition short-range displays, cathode-ray tubes with delay screens are totally unsuitable for use on small ships for manoeuvring or docking purposes. Tests on the *Elettra* brought this out very clearly. When manoeuvring in the Thames, rounding bends, etc., the displacement of the picture when turning, superimposing a new picture from a different aspect (owing to the turning of the vessel) upon a completely different background which was fading out, spoilt the definition of the picture when it was most needed. This effect is accentuated by the fact that the f.m. system gives a very bright display. It would seem that a completely different form of display is required, and this problem requires very careful thought. If it were solved, it should be possible to develop short-range high-definition f.m. radar apparatus which would be of the greatest use for navigation in restricted waters under fog conditions and for docking purposes.

The circuits of such apparatus would, with the exception of the microwave ones, be readily adaptable to transistor use, thus making possible a compact set with low power consumption in which the scanner could be placed high to obtain very clear short-range pictures.

Mr. C. F. Wilkinson: It is stated that a linearized detector is used in the receiver. In this type of receiver there would be some advantage in using a coherent detector instead of a linear or a square-law one. I estimate that, for a signal/noise ratio at the input of 2, an improvement of about 5 dB would be obtained at the output with a coherent detector. I notice that there is considerable noise on the 9-mile range (at least, I assume that it is noise) and a coherent detector should effect an improvement.

A number of pulse radar equipments operating on 1 000 yd range scales have resolutions no better than 1/30 and seem quite acceptable. Is a resolution very much better than this really required? What is the author's aim in this respect?

Mr. A. P. L. Milwright: The disadvantage of this type of radar equipment appears to be its complexity for its very limited application, compared with the conventional pulse radar. The very good range discrimination achieved is obtained at the expense of limited range of view. There are pulse equipments in use at present which have a range discrimination of approximately 12 yd when used with range scale of $\frac{1}{2}$ mile. To obtain this discrimination with the author's equipment, with a resolution of 1/30, would give a range of view ahead of only some 360 yd. I cannot imagine any master of a merchant ship attempting to navigate a narrow channel such as the Thames with so limited a field of view ahead, particularly when the data rate is so low (10 r.p.m., i.e. 6 sec per 'paint'). Even the low figure of 10 knots as the closing speed of two ships means that they have moved 100 ft, or 10% of the detection range, between scans. If the resolution is limited to these figures, I do not think that the equipment described will have any application for navigation.

However, where docking is concerned and very much lower speeds are involved, the much higher discrimination of which the f.m. radar is capable will be of great value. In these circumstances the radar might be sited ashore. I do not think that it would be attractive to the mariner to carry a complex set of this kind only for docking.

With pulse radar equipments the fractional resolution improves as the range scale increases, and at a 25-mile range scale the resolution approaches the spot resolution of the cathode-ray tube, which is about 1/200. What does the author think is likely to be achieved with improved values, and will the resolution ever approach 1/200? If it would mean an increase of frequency modulation or range of frequency tuning there may be trouble in getting an allocation of frequencies in the spectrum.

Mr. Rust mentions blurring of the picture owing to the turning of the ship; this need not arise if the radar picture is stabilized in azimuth from a transmitting compass. All conventional British pulse radar equipments are capable of being so stabilized.

Mr. P. S. Brandon: It may be of interest to compare the author's equipment with a parallel equipment developed for harbour approach. To increase the scan rate the aerial scan was limited to any selected 90°. With a repetition rate of 500 c/s to avoid possible range ambiguities, a 2° beam, and examining targets three or four times, a scan rate of 8 times per minute was achieved. A 1/20 fractional range resolution was obtained using beat notes between 0 and 50 kc/s. The panoramic receiver had a 500 c/s bandwidth and scanned 20 times a second. The display tube was 6 in, being equivalent to a normal 12 in display.

If the receiver was switched off for about 40% of the time, near-target spoking was avoided. Although gain modulation might be expected to degrade the range resolution, it was often found possible to counteract this by using larger effective frequency deviations.

The system used a separate local oscillator locked to 30 Mc/s from the transmitter. We took this on the Humber at 6.0 a.m. As we left the docks and went to the river centre, the temperature drop gave a differential drift of more than 7 Mc/s, with consequent breaking of the lock and difficulty of resetting. This led us to develop an automatic following system similar to that described. Because of gain modulation, the beat-note spectrum may be spread to about four times the modulation frequency; only 30 analysers are required in a multi-gate system equivalent to the radar described in the paper.

The ideal gain-modulation shape, from the point of view of getting a clean signal back from each target, is Gaussian error, but this does not fulfil the requirement to have the receiver dead during the flyback of the sawtooth modulation. A good compromise is a $\cos^2$ response, very much like that shown in Fig. 1.

The panoramic receiver also ideally requires a Gaussian response shape, and this can best be approximated by a series of single tuned circuits in cascade, coupled by valves. In many cases we had to compromise and have circuit pairs loosely coupled in the anodes of each valve. To cope with the large range of echo size, 4–6 effective circuits in cascade, six probably being the best, are needed.

Mr. G. R. Nicoll: The great difficulty with f.m. radar equipment is that of wasted information. In an equipment giving a fractional resolution of 1/30 and using only one filter, 29/30 of the transmitter power is wasted. In a pulse radar equipment effectively all the power is utilized. Therefore for f.m. radar to compete with pulse radar it seems to be necessary to use a full bank of filters. Does any new development in filter technique, using transistors for example, hold out promise of economical banks of filters? Is there any loss of signal/noise ratio in using what appears from Fig. 1 to be a sinusoidal modulation of the receiver gain? Is the modulation, in fact, sinusoidal?

One argument raised against pulse radar equipment for very-short-range working is paralysis of the receiver by the transmitter pulse. This is not necessarily a difficulty and, of course, will scarcely arise if separate transmitting and receiving aerials are used, as is often the case in f.m. equipments.

It is stated in the paper that velocity information is not normally required in the merchant service. Would it be desirable to have this information for abnormal situations, even at the expense of some additional complexity in the equipment?

Mr. J. Croney: I think it is true to say that any f.m. radar

quipment with a linear deviation of 100 Mc/s can give the same nge resolution as a pulse radar equipment using a 0·01 microsec ulse and a receiver bandwidth of 100 Mc/s. The generation of ch short pulses is quite difficult, and although recent develop- ents in amplifying valves have eased the production of the ide-bandwidth receiver, it would still be a fairly difficult and ostly process. Since this receiver problem does not occur in e f.m. case, what degree of linearity would be required of the 00 Mc/s frequency deviation in order to realize the ultimate efinition of the f.m. radar equipment? In other words, is it racticable to make the frequency deviation linear enough?

In those multi-range pulse radar equipments of which I have xperience, when I have selected a very short range the picture rightness has usually been disappointingly low. It seems a ositive advantage to f.m. radar that the full brightness is main- ained on all range scales.

To what extent does the performance of an f.m. radar equip- ent suffer from a pulse-radar environment? I feel that an f.m. adar equipment may pick up serious interference from numbers f pulse radar equipments spread throughout its deviation sweep. here are certainly circuits which can minimize this trouble, ut what is the author's experience?

It has been suggested that gain modulation has the effect of emoving the phase coherence between one sweep and the next, ut I still feel that in the f.m. equipment one preserves some hase coherence from one sweep to the next, and this could give ome of the advantage of coherent detection. Is an improved ignal/noise ratio possible from this phase coherence?

Dr. R. L. Smith-Rose: I remember vividly the introduction of .m. techniques by Appleton and his co-workers over 30 years go, and the original experiments carried out to demonstrate the xistence, and to measure the height, of the ionosphere. The echnique soon gave way to pulse modulation for this purpose, nd this has held sway ever since; but we were dealing with the neasurement of much larger distances than those with which the resent paper is concerned. I remember also that, after the ulse technique was introduced for radar, Appleton periodically xpressed the opinion that a fair opportunity had not been given o the possibilities of frequency modulation for radar purposes.

While most of the discussion is about ships, there might be a nore obvious application for harbour-entrance work with the quipment mounted on shore. What does the author think of the merits of the f.m. technique in a place such as the entrance to the Mersey, where I believe that a pulse-radar equipment is used to survey the position over a quite limited range and sector? This would appear to be one of the cases where a real comparison could be made, giving a figure of merit for each of the two systems.

Dr. E. Eastwood: On the question of a multi-gate receiver, Mr. Brandon could have referred to the fact that he has been responsible, with Mr. Rust, for a radar in which 400 filters were utilized quite satisfactorily. This work has fully established the practicability of a multi-filter f.m. system.

In the discussion there has been comparison of the f.m. and pulse systems. Such comparisons are inevitable if the f.m. equipment seeks only to emulate a pulse equipment, but I suggest that we should regard the two techniques, not as alternatives, but as complementary in the task of determining the position and velocity of a target.

Problems associated with the frequency sweep of the f.m. radar unit are obviously best avoided by abandoning the sweep and using the Doppler information derived from the resulting c.w. system. This can be reduced to an extremely simple device capable of supplying azimuth and velocity information on moving targets, i.e. targets possessing a velocity relative to the equipment. This information is sufficient in itself for many important applications; it can also complement the information derived from a pulse system without involving the difficulties and complexities of a high-data-rate f.m. ranging radar system.

Mr. M. Morgan (*communicated*): Adding to Mr. Brandon's account of a harbour approach radar, it may be of interest to know that its first range was 10–100 yd. When the equipment was tried out on the roof of the north wing of our laboratories near Chelmsford, one could pick out the individual ventilators in the rows on top of the northern lights, the central-heating chimney showed up well, and beyond the edge of the building two masts and a piece of rope and corner reflector, hanging between them, were easily identified. The general outline of the main building and the garage at the edge of the grounds and some of the surrounding fencing were clearly shown. Some trouble was experienced from overloading produced by very near targets, and showing up as enlarged areas on the p.p.i. This effect was largely eliminated by putting a suitable filter in the beat-note circuits to counteract the inverse fourth-power range law.

THE AUTHOR'S REPLY TO THE ABOVE DISCUSSION

Mr. D. N. Keep (*in reply*): I agree with Mr. Rust that gain nodulation has a further important function on the very-short- ange scales, where frequency modulation is likely to be very ion-linear at the extremes of the sweep. In general, however, here the linearity of sweep is adequate the effect of gain nodulation is to widen the spectrum of the target to some extent; he consequent loss in signal/noise ratio is small.

It is undoubtedly true that better short-range pictures can be btained with similar transmitting valves, but only, in my pinion, under artificial conditions which are very favourable to he radar equipment. Very great care was used in setting up the ain modulation for the range scale in Fig. 4, and the differences etween targets (*a*) and (*e*) are readily explained, since they are ifferently orientated and at very different ranges. When xamined, on an A-display, target (*e*) showed a considerably vider spectrum than target (*a*).

For the longer-range scales the gain-modulation conditions sed approximated to those mentioned by Mr. Brandon.

I agree with Mr. Milwright that the present equipment is nadequate for operational use. An equipment using four arallel sweeping gates could be constructed to give a range resolution of 1 in 120; with a range resolution of 5 yd, the minimum range scale would be 600 yd. In order to make the best use of the bright picture obtained it would be necessary to stabilize the display, as he points out.

It seems likely, however, that a shore-based docking radar equipment would be a more attractive proposition, and under these circumstances a multi-gate receiver might be of practical use, since minimum size and cost are no longer of major importance.

The siting of the equipment would have to be very good, however, in order to obtain adequate azimuthal resolution, and blanking by large obstacles might prevent the use of very short ranges. As Dr. Smith-Rose suggests, the Mersey would be a good place for the operational assessment of such an equipment.

For a range resolution of 1 in 200, a bank of 200 filters would be required, and with a repetition rate of about 1 kc/s the beat notes would fall within the operating ranges of transistors; the receiver could therefore be reasonably compact. The aerial rotation rate would be of the order of 4 r.p.m. Higher data rates could be achieved with higher repetition rates and conventional valve amplifiers. The linearity of sweep required for a

resolution of 1 in n is approximately $1/n$; hence this radar equipment would require a sweep linearity of $\frac{1}{2}\%$. It seems likely that this could be achieved with the ferrite techniques mentioned by Mr. Rust, even over as wide a sweep as 100 Mc/s.

Much interference was experienced with the set described owing to very-high-power pulse sets operating in the vicinity (within 100 yd). However, under operational conditions this problem should scarcely exist, although it might occur over limited sectors when the density of shipping was high.

Both pulse-cancellation and pulse-clipping circuits have been used with a certain amount of success, as Mr. Croney suggests, but this is naturally dependent upon the characteristics of both the equipment and the interference.

No coherence exists between the bursts of echo energy from sweep to sweep unless there is an integral relation between the beat frequency and the modulation repetition rate. The latter condition has been exploited in an equipment for measuring the height of the ionosphere (see Reference 6). Unfortunately, the method does not appear to be practicable in high-data-rat radar equipments, and where the target is moving the integratio must take place in a Doppler filter.

On the general question of operational requirements, no wor has been done which gives definitively the limits required fo very short-range working. There is no doubt in my mind that resolution of 30 yd is completely inadequate; 12 yd is approachin the sort of thing required, but I feel that a resolution of abou 5 yd is really necessary.

The extraction of velocity information would increase th complexity of the apparatus and would be mainly indicative o collision courses. The information obtained would require ver careful assessment and might even be misleading. I do no think, therefore, that its use would be justified. I agree wit Dr. Eastwood that there are many important application requiring azimuthal and velocity information only, but I do no think that navigational use in the mercantile marine is one o these.

621.395.11

Paper No. 2082 R
July 1956

CHANGE OF PHASE WITH DISTANCE OF A LOW-FREQUENCY GROUND WAVE PROPAGATED ACROSS A COAST-LINE

By B. G. PRESSEY, M.Sc.(Eng.), Ph.D., Member, G. E. ASHWELL, B.Sc., and C. S. FOWLER

(*The paper was first received 7th December, 1955, and in revised form 17th February, 1956.*)

SUMMARY

Observations of the change of phase with distance have been made on a frequency of 127·5 kc/s along a number of paths radiating from a transmitter near Lewes and crossing the south coast between Pevensey and Littlehampton. The nature of the ground adjacent to the coast, the angle of crossing the coast-line and the lengths of the land and sea sections covered varied from path to path; the greatest distance covered out to sea was 22 km. Measurements were also made over paths at right angles to the radials, and the phase changes in the area off Worthing, where the paths were tangential to the coast, were examined in detail.

The results confirm the existence of a phase-recovery effect which, as theoretical considerations have shown, should be experienced by a wave passing from low-conductivity ground to sea water and which was indicated by previous measurements over geological boundaries on land. The detail of the measurements at sea shows also that, in addition to this general behaviour of the phase, there are superimposed systematic variations whose magnitudes decay from about 4° of phase near the coast to a negligible amount at 6λ out to sea and on some paths are comparable to the recovery. A very marked phase disturbance within λ/2 of the coast on the landward side is also evident; it is similar to that previously observed over geological boundaries on land.

(1) INTRODUCTION

This investigation is a continuation of the work carried out in previous years on the phase change with distance of low-frequency waves propagated over inhomogeneous ground.[1] One of the main features of the results of this work was the nature of the phase change at a boundary between soils of different conductivities. It was considered that measurements over a coastal boundary would enable this change to be studied under more ideal conditions: a well-defined boundary would be provided, and, with the transmitter on land, continuous phase observations could readily be made beyond the boundary. Again, the freedom of movement and freedom from local site errors of a sea-borne receiver station would greatly facilitate the exploration of the phase pattern over a wide area and the determination of the phase changes normal to the path of propagation.

The measurements were made on the l.f. waves transmitted from the green slave station of the English Decca Navigator chain on a frequency of 127·5 kc/s. This station is situated near Lewes, Sussex, and the main measurement paths were radials from this point crossing the coast between Brighton and Pevensey. Other radial paths running tangential to the coast and over land and paths lying normal to the radial ones were also included, so that a comprehensive survey of the phase pattern over an area might be obtained.

The phase-measuring equipment used in this investigation is described briefly in Section 2, but a complete description is given elsewhere.[2]

In addition to the phase measurements the observations at sea included readings on a standard Decca receiver and bearing measurements with a l.f. direction-finder. The directional measurements are described elsewhere.[3]

(2) METHOD OF MEASUREMENT

(2.1) General Principles

The general principles underlying the method of measurement are illustrated by Fig. 1. The low-frequency transmitter situated many wavelengths inland from the coast provides c.w. signals

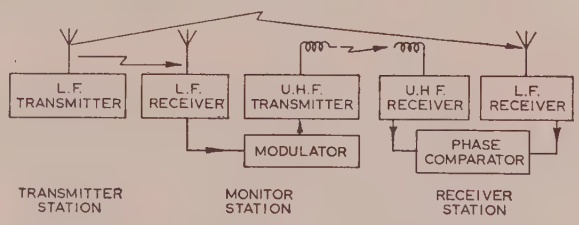


Fig. 1.—General arrangement of measuring equipment.

which are received at a monitor station at or near the coast and at a mobile receiver station on land or at sea. At the monitor station the l.f. signal is used to generate modulating pulses for a u.h.f. transmitter at a repetition frequency which is a simple fraction of the frequency of the l.f. signal. A monitoring system ensures that the time of occurrence of the pulses bears a constant relation to the phase of the incoming l.f. signal. At the receiver station the u.h.f. signal is demodulated and the resultant pulses are applied to a receiver from which is obtained a sinusoidal reference signal. The phase difference between the l.f. signal and this reference signal is measured by means of a phase comparator. Apparatus is provided for checking the constancy of the phase shift through the equipment before each measurement.

The three stations are normally located on the same straight line, and as the receiving station moves out along this line the change in the measured phase difference will give the actual phase change of the l.f. wave relative to the value it would have if its phase velocity were equal to the group velocity of the u.h.f. wave. Since the velocity of the u.h.f. wave is uniform with distance, is unaffected by the nature of the ground and its value is readily calculable, the total phase change of the l.f. wave can be determined.

In the calculation of the velocity of the u.h.f. signal it is assumed that the transmission is by the space wave and that the dispersion of the atmosphere is negligible. Under these circumstances the group velocity will be equal to the phase velocity, the value of which will be determined solely by the refractive index of the air. Under the atmospheric conditions likely to be met with over sea it is estimated that the velocity will vary from about 299 670 to 299 700 km/sec. This variation is of the same order as the expected accuracy of measurement, but, if required, a more exact value for the velocity can be calculated from the measured atmospheric constants.

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.
The paper is an official communication from the Radio Research Station, Department of Scientific and Industrial Research.

(2.2) The Receiving Stations

The receiving equipment was installed for the sea trials in the wheelhouse of a 72 ft survey motor launch lent by the Admiralty Hydrographic Department. The u.h.f. aerial was mounted on a 20 ft rotatable mast strapped to the side of the bridge, the short l.f. aerial being supported by the same mast. The equipment was earthed via the bonding system of the ship to copper plates on the hull.

The monitor equipment and the receiving equipment, when used for the land measurements, were fitted in 5-ton vans with aerial systems similar to those described above mounted on their roofs.

(2.3) Observational Procedure

(2.3.1) At Sea.

Measurements of phase and the ship's position were made simultaneously at 1–2 min intervals. The phase indicator was kept under continuous observation, and by recording the average reading over a 5–10 sec period, errors due to the rapid fluctuations by atmospheric noise were eliminated. The zero setting of the equipment was checked at frequent intervals.

The ship's position was determined by sextant observations on known landmarks. A running plot of the position was made and the ship's course was maintained as closely as possible on the desired line.

(2.3.2) On Land.

Two or three phase measurements over a period of 5 min were taken at each land site. The positions of the sites were fixed by reference to 6 in Ordnance Survey maps. The spacings between the sites varied from 50 m to 2 km according to the detail of measurement required or to the availability of suitable sites. Measurements were made at one or two check points at the start and end of each day's work in order to correlate the various series of observations.

(2.4) Accuracy of Measurement

The factors which affected the accuracy of measurement were the instrumental stability, the earthing conditions at receiver and monitor stations, the relative positions of the l.f. and u.h.f. aerials at the receiver station and the deviations of the measurement position from the radial path.

Previous tests had shown that the phase fluctuations which could be attributed to instability in the equipment had a standard deviation of less than 0·3° (at the transmission frequency of 127·5 kc/s), and the investigations showed that this was much less than errors due to other causes. The earthing conditions at the monitor station and on the receiving van can produce changes in the phase of the aerial circuit impedance. In order to minimize these effects the monitor station was earthed through one or more steel rods and the receiving van had no direct earth connection. The resulting phase errors for a particular series of measurement were small, but when the monitor station was moved from one site to another or returned to the same site on the following day the variations observed at the check points had a standard deviation of 1·8°. The effect of these variations was largely eliminated by the application of corrections derived from the check-point readings.

Errors due to the changes in the relative positions of the l.f. and u.h.f. aerials when the ship was turned were measured and found to be less than 0·5°, which is in agreement with the calculated value of 0·3°. Although the results showed a tendency for the readings taken on the outward runs to be about 0·3° higher than those taken on the inward runs, no correction was applied since it was of about the same order as the random errors. In the case of the receiving van the phase change on turning about was greater, 1·5°, but at the majority of sites it was possible to orientate the van in the same direction relative to the transmitter and so make any correction unnecessary.

Deviations of the measurement positions from the radial paths necessitated the application of corrections to the majority of the phase readings. This was particularly important on some of the cross-paths, where the corrections exceeded the expected phase changes. The need for such corrections arises from the fact that at points off the radial from the transmitter through the monitor the difference between the l.f. and u.h.f. transmission paths is not constant and equal to the distance between the transmitter and the monitor. This excess path difference can be readily calculated and the phase correction determined. For a given distance, the greater the deviation of the measurement positions from the radial, the more accurately must the magnitude of that deviation be known. For example, at 5 km from the monitor a deviation of the order of 500 m must be known to an accuracy of ±30 m if the corrected phase reading is to be accurate to 0·5°; but if the bearing of the receiving station from the monitor can be maintained to within 1° of the radial, the phase error will be less than 0·5° for distances up to 22 km from the monitor and no correction is required. On the radial runs at sea, navigation to this accuracy and better was achieved over practically the whole of each course, the main exceptions being in those regions where the ship was within a kilometre of the monitor. It is considered that in all cases the ship's position was known to an accuracy of ±20 m or better at each fix and the phase error due to uncertainty of position was normally less than ±0·2°. On land the position of the receiving van could be fixed to an accuracy of about ±3 m. It is estimated that the overall errors of measurement had a standard deviation of 0·5° at sea and 1° on land.

The possibility of errors arising from instability of the reference phase due to multi-path transmission of the u.h.f. signal was considered and formed the subject of a preliminary investigation. It was found that large errors, up to 5° of phase, could be caused by ships crossing or approaching the path between the monitor and receiver stations. In the present investigation, however, such circumstances did not arise. On the other hand, when the u.h.f. path lay partly or wholly over land there was still the possibility of disturbances to the reference phase due to the reception of secondary signals which had been reflected from buildings, hills etc. In practice, however, no evidence was found during the measurements of either amplitude or phase variation of the reference signal which might be attributed to this cause.

(3) MEASUREMENT PATHS

All the paths over which phase measurements were made are shown in Fig. 2. These paths were chosen so as to cross the coast at various angles and to cover the widest range of conductivity ratios between land and sea that are available in the area.

The Worthing area, in which the l.f. transmission path becomes tangential to the coast, was examined in detail by phase measurements made on the radial paths 6, 7, 8, the cross-paths 10, 11, 12 and two unnumbered paths parallel to path 8 and spaced approximately 1·5 and 3 km away. The phase changes in the immediate vicinity of path 5 were also explored in detail by making a succession of transverse runs (cross-path 13) and a number of runs along paths radiating from the monitor station and lying within a few degrees of the main path.

Land measurements were made along sections of paths 1, 3, 4, 8, 9, the extensions of cross-paths 10, 11, 12, and at other points in the area between radials 8 and 9. A further land run (cross-path 14) was made along the coastal road between Balsdean and the eastern end of Peacehaven, the monitor site used being situated at High Dole, 2 km inland from the mid-point of the path.

On some paths two monitor sites (indicated by A and B in

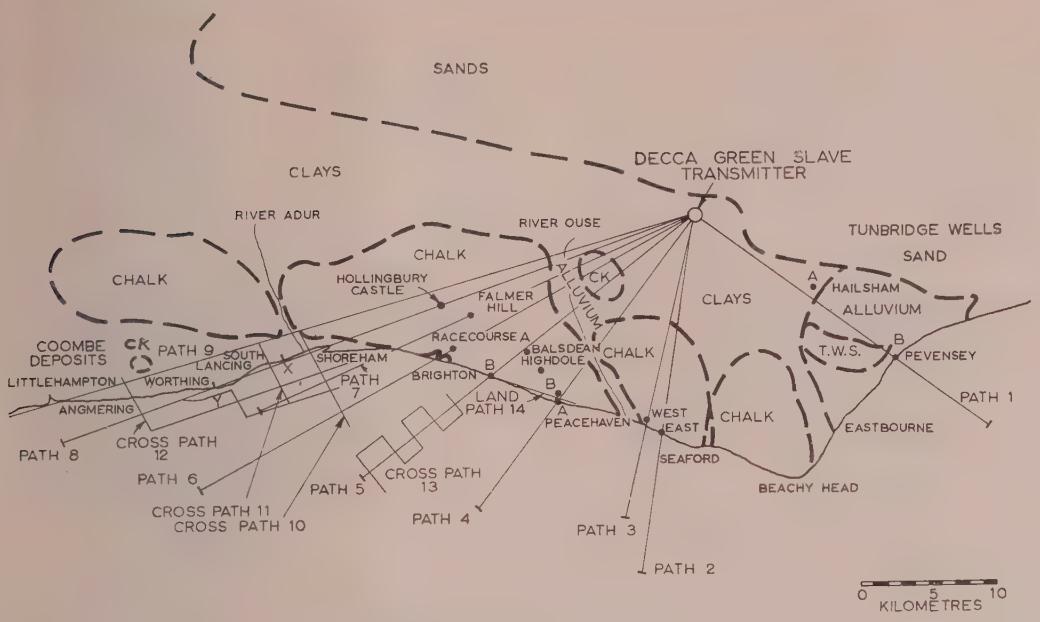


Fig. 2.—Sketch map showing measurement paths.

—— Paths.
● ● Monitor sites.

Fig. 2) were used, one on the coast and the other further inland. This change in the position of the monitor station was necessary in order to increase the height of the u.h.f. transmitter, to obtain a clear transmission path or to extend the measurements inland.

The character of the land–sea boundary varied between the one extreme at Pevensey Bay, where the land is low lying and of high conductivity, and that at Peacehaven and Balsdean, where the transition from the medium conductivity of the chalk to the high conductivity of the sea water takes place over the edge of 100 ft cliffs. Between Shoreham Harbour and Worthing, path 8 runs along the coast on the landward side, but in the analysis of the results it has also been treated as though this section were entirely over sea. The path lengths were limited in most cases at sea by lack of adequate visibility for accurate position fixing.

(4) EXPERIMENTAL RESULTS

The measured phase changes along the radial paths are plotted in Figs. 3–10. On most paths measurements were made on both the outward and inward runs over the sea section, and in some cases additional check runs were made over part of the section: all points are included in the Figures. In Fig. 7 (path 5) the measurements made along one of the secondary radials (path 5A) are also included.

The phase changes along cross-path 10 are shown in Fig. 11. The results along cross-paths 11 and 12 were of the same general nature in that they showed a steady increase of phase as the coast-line was approached with a more rapid increase over land.

The measurements in the neighbourhood of path 5 are shown in Fig. 12 plotted on isometric axes.

The measurements on the sea section of the Peacehaven path have been replotted on an extended scale in Fig. 13, together with the measurements made on land in this neighbourhood. The two land runs were not made along the path: that based on the Peacehaven monitor site B followed a line at 22° to the path, and that based on the High Dole monitor followed the coast road. But for the purpose of investigating the phase changes near the

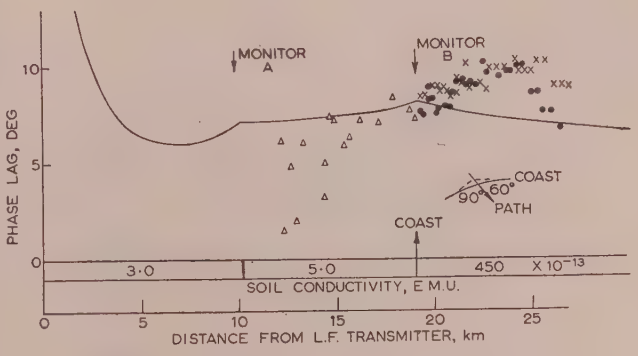


Fig. 3.—Phase change along path 1 (Pevensey).

—— Theoretical.
× × Outward sea run.
● ● Inward sea run.
△ △ Land run.

coast-line all these results have been plotted at the appropriate distance from the coast as though all the sites had been located on the radial path.

The sites used for the land measurements along paths 8 and 9 were often considerable distances off the paths, owing to the difficulties of finding suitable locations at the required distances, and a particular site could not always be associated with a particular radial. The results of measurements along the two paths have therefore been plotted together in Fig. 10, but some indication of the positions of the sites relative to each path is given by the symbols used.

In plotting the phase values measured at sea and on land on the same graph, due allowance must be made for the discrepancy due to the differences between the aerial installation on the launch and that on the van, since the receiving-equipment monitoring system does not cover the phase shifts in the aerial circuits. The installations differed in respect of the relative positions of the

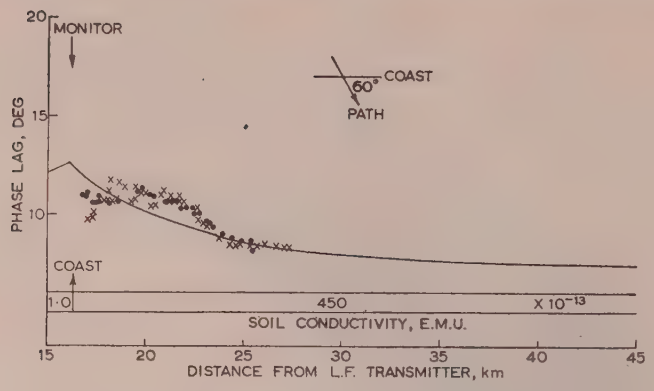


Fig. 4.—Phase change along path 2 (Seaford East).

—— Theoretical.
× × Outward run.
● ● Inward run.

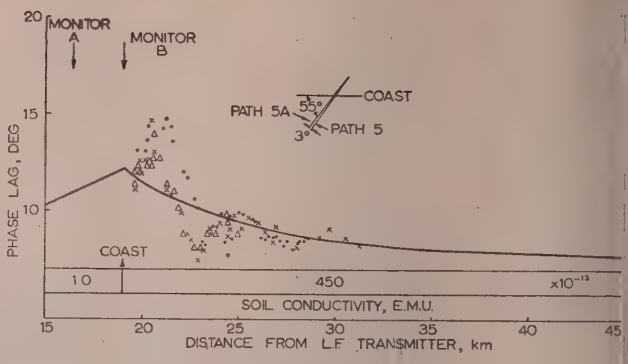


Fig. 7.—Phase change along path 5 (Balsdean).

—— Theoretical.
× × Inward run along path 5A; monitor at A.
△ △ Inward run along path 5A; monitor at B.
● ● Inward run along path 5; monitor at A.

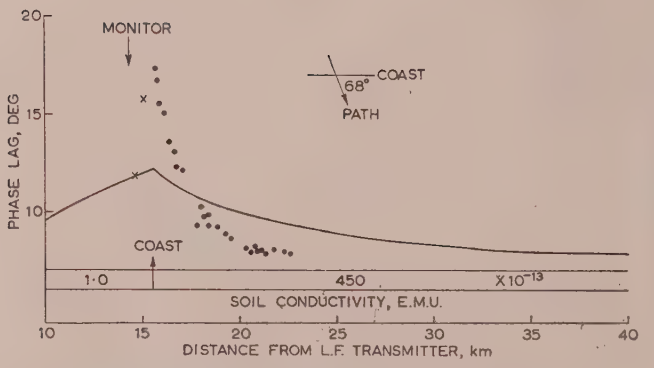


Fig. 5.—Phase change along path 3 (Seaford West).

—— Theoretical.
● ● Outward run.
× × Land run.

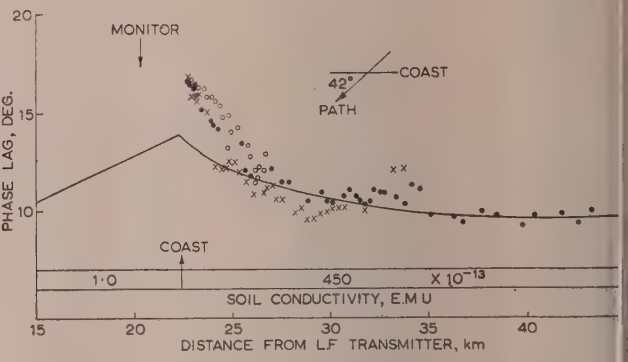


Fig. 8.—Phase change along path 6 (Brighton race-course).

—— Theoretical.
× × Outward run.
● ○ Inward runs.

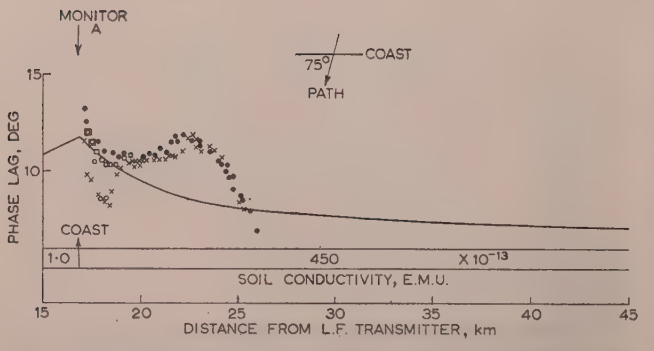


Fig. 6.—Phase change along path 4 (Peacehaven).

—— Theoretical.
○ × Outward runs.
□ ● Inward runs.

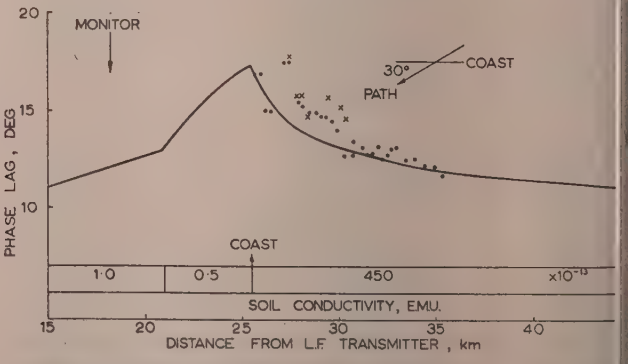


Fig. 9.—Phase change along path 7 (Falmer Hill).

—— Theoretical.
× × Outward run.
● ● Inward run.

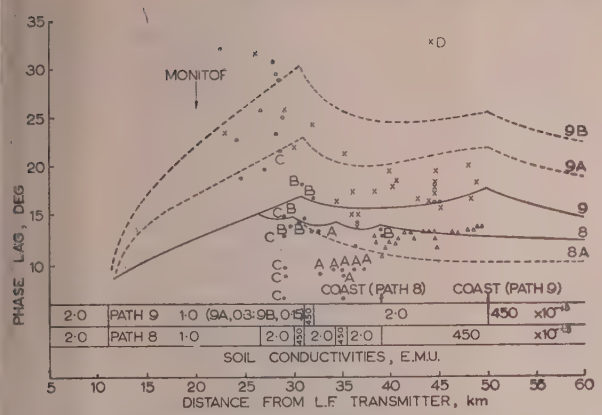


Fig. 10.—Phase change along paths 8 and 9 (Hollingbury Castle).

Theoretical.
Points within 1° of path 9.
Points between paths 8 and 9.
Points within 1° of path 8.
Outward sea run.

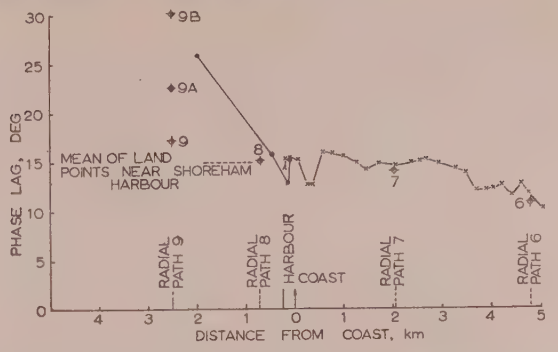


Fig. 11.—Phase change along cross-path 10 (Shoreham Harbour).

Land measurements.
Sea measurements.
Theoretical values.

.f. and u.h.f. aerials, the length of feeder from the u.h.f. aerial and the earthing conditions. Since it was not possible to calculate the phase correction, it was determined by inspection of measurements made in three areas where land and sea sites lay within a few hundred metres of each other. One of these was Shoreham Harbour, where several readings were taken both in the harbour and on the quayside. The average difference in each area was 3° of phase, and this amount was added to all the readings taken at sea so as to bring them in line with the land measurements.

All phase readings were relative to the phase of the reference signal at the monitor station, so that in order to relate one set of readings to another it was necessary to know the relative phases of the reference signals at the monitor positions. Where the monitor station was moved from one site to another on the same path, as on paths 1 and 4, the resultant change in the reference phase was measured directly at the receiving station. For paths 6, 7, 8 and 9 the cross-runs provided links between the readings taken on these paths and enabled the relative phases of the reference signals at three monitor sites to be determined. The alignment of the readings along the other paths was obtained with the aid of the theoretical curves of phase change with distance as explained in the next Section.

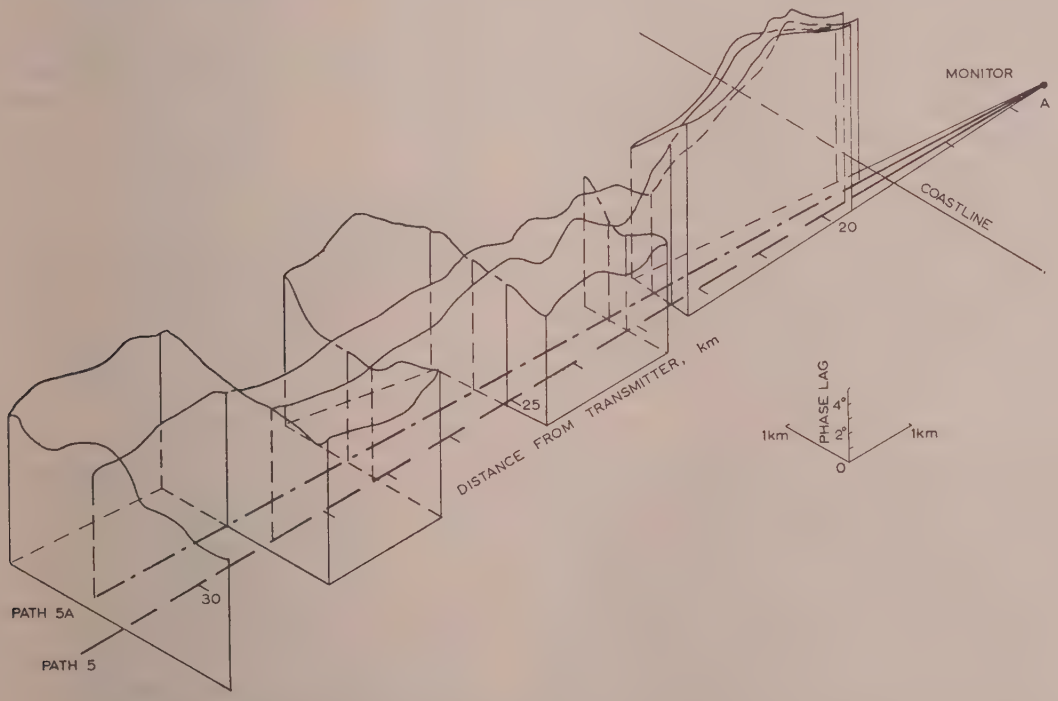


Fig. 12.—Isometric plot of phase variations over sea in the neighbourhood of path 5.

(5) THEORETICAL PHASE CHANGE

The method used for constructing the theoretical curves of phase change with distance shown in Figs. 3–10 was that adopted in earlier investigations over land paths.[1] It consisted essentially of a process of building up a composite phase curve from the theoretical curves for homogeneous ground which were appro-

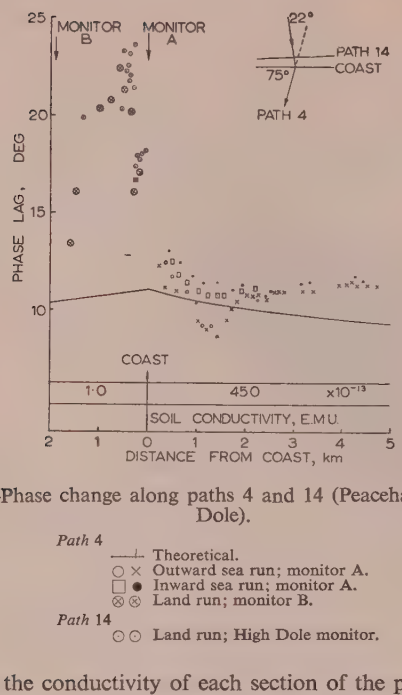


Fig. 13.—Phase change along paths 4 and 14 (Peacehaven and High Dole).

Path 4
—|— Theoretical.
○ × Outward sea run; monitor A.
□ ● Inward sea run; monitor A.
⊗ ⊗ Land run; monitor B.
Path 14
○ ⊙ Land run; High Dole monitor.

priate to the conductivity of each section of the path. At that time the method had no rigorous theoretical basis and relied for its justification on the good agreement found to exist between calculated and measured values. Since then, however, a mathematical analysis of the propagation of ground waves over a mixed path[4] has given results in close agreement with those obtained by this method, thus increasing our confidence in it. The conductivity of the ground was estimated from a knowledge of its geological structure using the experience gained in the previous work.

It is to be noted that the phase value obtained by this method is the amount by which the phase change over the ground exceeds that in air, and is therefore the same quantity as that measured in the investigation. It is calculated that for the atmospheric conditions obtaining at the time of the measurements the velocity in air was 299 690 $\pm$ 10 km/s. In Fig. 11, showing the phase changes along a cross-path, the only theoretical values given are those corresponding to the intersections with the radial paths.

The datum phase level for the theoretical curves is the phase at the l.f. transmitter, whereas for the measured phase values it is the phase of the reference signal at the monitor station. The theoretical curves give a value for the phase of the l.f. signal at the monitor site, and if the phase shift through the equipment from the l.f. receiving aerial to the u.h.f. transmitting aerial is known, the phase of the reference signal can be calculated and the measured phase changes adjusted accordingly so as to line up with the theoretical curve. Although the phase shift through the equipment can be, and was, maintained at a constant value, the phase shift in the l.f receiving aerial is unknown and varies with the earthing conditions, as pointed out in Section 2.4. Thus there is an uncertainty of about 3° in the adjustment of the measured phase by this method. An alternative method of adjustment is to fit the measured values to the theoretical curve near the end of the path, where, in general, the fluctuations are least. Although such a method is reasonable for the longer paths, it is not really satisfactory on the shorter ones. It was therefore decided to use both methods in making the adjustment and to adopt a compromise where they did not give the same result. In doing this, account was taken of the other data available, e.g. the links between two or more sets of readings provided by the transverse runs both at sea and on land, the check measurements made at specific land sites on different days and the expected variations of phase due to the earthing conditions.

(6) DISCUSSION OF THE RESULTS

(6.1) On the Radial Paths across the Coast-Line

Of the seven radial paths (1–7) which cross the coast-line the measurements on five (2, 3, 5, 6 and 7) show a definite decrease in phase lag after the wave has crossed the coast-line (see Figs. 3–9). Moreover, the phase change follows the theoretical curve reasonably closely: this is particularly noticeable on the longest path (6), which was long enough to reach the point at which the phase change had begun to settle down to a value characteristic of propagation over sea. The angle at which the paths crossed the coast varied from 30° on the eastern side of the normal to 60° on the western, but there appears to be no significant effect due to this variation.

On the other two paths (1 and 4) the agreement between the measurements and the theoretical values is not so good, although on path 4 there is a general tendency for the phase lag to decrease after the wave has crossed the coast. On path 1 there is an unexpected increase in phase lag beyond the coast-line, but it must be noted that in this case the theoretical phase disturbance at the coast is comparatively small and the difference between measured and theoretical values is approximately the same as for path 4.

On all paths the measured phases show a regular deviation from the theoretical curve over approximately the first 14 km of the sea section of the path. The pattern of these deviations is shown more clearly in Fig. 14, in which the smoothed mean curve

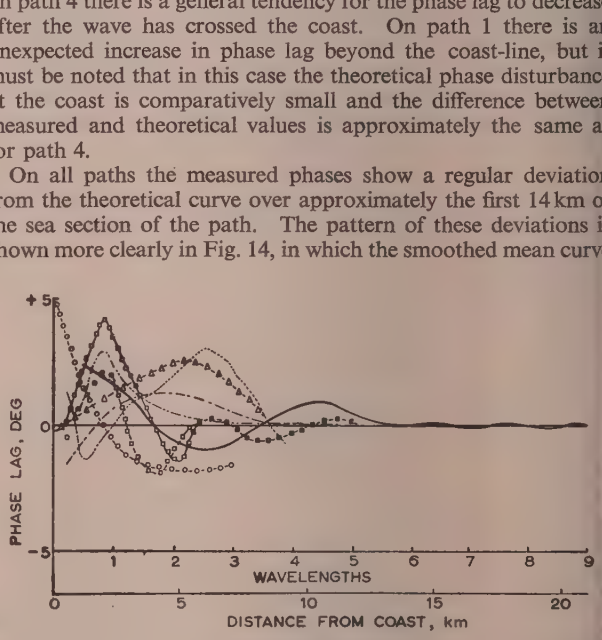


Fig. 14.—Deviation of mean measured curves from the theoretical phase change.

—△—△— Path 1.
—·—·— Path 2.
—○—○— Path 3.
· · · · · Path 4.
—□—□— Path 5.
—■—■— Path 5A.
——— Path 6.
—·—·—·— Path 7.

through the measured points is plotted relative to the theoretical curve for all the radial paths. The standard deviation of the points from the mean curve is about 0·5° in all cases and corresponds to the estimated accuracy of measurement. There seems to be little similarity between these curves; the only feature which they have in common is the manner in which their amplitudes decay to a negligible value in a distance of about 6λ from the coast. The measurements made in the neighbourhood of path 5 showed that similar variations of phase occur across the radia

aths. The results of these measurements, given in the isometric lot in Fig. 12, show clearly the undulating nature of the phase urface. This surface becomes smoother as the distance from the oast increases and its general slope follows that predicted by the heory. The cause of these undulations has not been established, ut it seems most probable that they are due to interference by number of secondary waves which have been reradiated from oints along the coast or very close to it. The location of the ources of reradiation is indicated by the relatively rapid decrease amplitude of the undulations with distance from the coast approximately according to an inverse-distance law), and the iversity in the phases of the secondary waves, which is assumed o be the cause of the random nature of the undulations, can be ccounted for by the irregularities of the coast-line and of the round behind it.

The number of measuring sites on land and the lengths of path hich they covered were, with one exception, too small to make omparison with the theoretical curves worth while. On path 1 Fig. 3), however, the measurements continued inland a distance f 8 km, but little agreement with the theoretical phase values as obtained. The measured points have the wide scatter bserved over other land paths (see Fig. 10). The measuring ites were located at various distances up to 1·3 km off the path, nd the variations in phase from one to the other indicate the xistence of a complex phase pattern for which the hill of Tunridge Wells sand just to the west of the path (marked T.W.S. n Fig. 2) may be mainly responsible.

(6.2) At the Coast-Line

The measurements made on land in the Peacehaven area and lotted in Fig. 13 show the phase disturbance which takes place s the wave approaches the immediate vicinity of the coast-line. he rapid increase in phase lag as the wave approaches the oundary, followed by a still more rapid decrease across the oundary, corresponds closely with the changes previously found o occur across geological boundaries on land.[2] In that case, owever, there was a more pronounced dip immediately beyond he boundary than is shown in Fig. 13. There is evidence of a imilar phase change on path 3 (Fig. 5), although the amplitude f the increase is not so great nor is the fall so rapid as in the eacehaven case. On path 1, however, where the conductivity f the land is high, there was no such phase change.

The reason for this marked local disturbance is not clear, but t seems possible that, like the undulations over the sea, it is due o reradiation from the boundary, which in this region would ave a greater amplitude and a more rapid variation of phase ith distance.

(6.3) Over Land Paths

The results of the measurements on paths 8 and 9 (Fig. 10) re characterized by the large changes of phase from site to site roducing a wide scatter of the points about the theoretical urves. It is noticeable that the few sea measurements which ere made at the end of path 8 have a considerably smaller ariation. This scatter is considerably greater than any random ariations due to the measurement errors discussed in Section 2.4 nd must be attributed partly to the existence of a complex phase attern in the region and partly to local phase disturbances at the easuring site.

The wide scatter of the experimental points renders it mpossible to do more than to make a very general comparison ith the theoretical curves. On path 8 the points (dots) lie ostly well below the full-line curve and closer to the broken urve 8A, which is based on the assumption that the path from horeham Harbour onwards is effectively over sea. Two groups f points on this path are of special interest. The first consists of the points marked A, which, except for the first, are 4°–5° lower than those on the adjacent sections of the path marked B. The measurements which these points represent were made along that section of the coast road between Shoreham and Worthing (X to Y in Fig. 2) where the path runs tangential to the coast-line. The points marked C comprise the second group and represent the results of measurements at sites within an area of 400 m radius near the entrance to Shoreham Harbour. The large differences between them is indicative of the complexity of the phase pattern in such an area.

On the section of path 9 which lies over the Coombe Deposits (31–50 km from the transmitter) the measured values (shown by crosses) lie within a few degrees of the theoretical curve. On the chalk section of the path, however, the measured values are appreciably higher than the theoretical ones. This also applies to points (circles) which represent measurements taken on sites lying between paths 8 and 9, and it is clear that over the whole of this area the phase values are much higher than was expected. Two possible explanations have been considered. The first was that the actual conductivity of the chalk was lower than the estimated value. If a value of $0{\cdot}3 \times 10^{-13}$ e.m.u. is used in the determination of the theoretical curve (9A, Fig. 10) the discrepancy is reduced. If, however, better agreement is sought by using an even lower conductivity—which, incidentally, is difficult to justify on geological grounds—the agreement over the remaining section of the path is considerably worse, as curve 9B shows. The second explanation considered was that, since this section of the path lies over the South Downs and rises to a height of 500 ft above sea level, the velocity of propagation might be reduced by the diffraction of the waves over the hills. If, however, an approximate calculation is made assuming the ground to have an equivalent radius of curvature of 350 km, it is found that over this 20 km section of the path the total increase in the phase lag is not more than 3°. Thus, although these high phase values may be partially explained in the above ways, their main cause must be sought elsewhere, probably in the local phase disturbances due to irregularities of the ground or surface objects.

The exceptionally high phase value given by point D, 44 km along path 9, was measured at a site about 1½ km to the north of the path, and may be explained by the fact that the l.f. transmission path to this site lay over an isolated chalk hill of lower conductivity than the surrounding ground (see Fig. 2). Unfortunately there was no opportunity to make further measurements in this area to determine the extent of this apparent local phase disturbance.

(6.4) On the Cross-Paths

The phase changes along the cross-path (Fig. 11) show that, apart from a small local disturbance at the coast there is, in general, a smooth variation of phase as the propagation path approaches the tangent to the coast from the seaward side. To some extent this is to be expected, since, owing to the shape of the coast-line, the lengths of the land and sea sections of the propagation path also changed in a continuous manner as the tangent was approached. A more abrupt change in the length of the sea section might give rise to an entirely different phase change on such a cross-path, particularly if the conductivity of the land was lower than in the present case.

(7) CONCLUSION

The investigation has provided a quantity of detailed information on the phase changes of an l.f. ground wave in the neighbourhood of a coastal boundary. While the results help to confirm phenomena previously observed over geological boundaries on land, they have also demonstrated how complex those changes really are when studied in detail. Thus it follows that the con-

clusions drawn from the investigation must not be regarded as being of general application in all respects.

The conclusions may be summarized as follows:

(*a*) The change of phase with distance of a ground wave crossing the coast-line from the landward side shows a 'recovery effect', i.e. the rate of change of lag, relative to that over ground of perfect conductivity, reverses from positive to negative sign at the coast and then slowly returns towards the value appropriate to transmission over sea. This general behaviour is in accordance with that determined theoretically.

(*b*) In addition to this general change there are superimposed variations which are indicative of an undulating phase surface with amplitudes of up to 4° near the coast dying away to a negligible amount at about 14 km (6λ) out to sea. On some of the paths these variations are comparable in magnitude to the recovery effect. Although no definite evidence as to the cause of these undulations has been obtained, it is concluded that they are probably due to the presence of secondary waves reradiated from points along the coast-line or in the neighbourhood of it.

(*c*) Within $\lambda/2$ of the coast-line there is a pronounced phase change which cannot be accounted for by the simple theory developed hitherto, but which has the same characteristic form as that previously found to be associated with an abrupt geological boundary on land. A marked increase of phase occurs before the boundary and is followed by a sharp fall across the boundary itself.

(*d*) Although certain anomalies were observed when the propagation path was tangential to the coast, the general phase pattern in this area showed a smooth transition from sea to land. It is considered, however, that in other areas quite different results might be obtained.

Considering the practical application of the results to the operation at sea of such c.w. navigational aids as the Decca Navigator, we may conclude that at distances of 5λ or more from the coast the phase change may be determined theoretically by the method previously given[1] and the errors due to coastal effects may be calculated. Nearer to the coast, however, the undulations of the phase surface are appreciable and the prediction of errors is impracticable, although some indication of their possib magnitude can be obtained from the measurements.

(8) ACKNOWLEDGMENTS

The authors gratefully acknowledge the very valuable assistanc given by the Admiralty Hydrographer in placing the launch a their disposal and the help of the Admiralty Signal and Rada Establishment in obtaining that assistance. They are als indebted to Lieut. W. D. Stuart, R.N., the officer in command c the launch, for his ready co-operation and for making the larg number of position fixes required.

Mr. H. E. Brown was responsible for the operation of th monitor station.

The work described above was carried out as part of th programme of the Radio Research Board. The paper published by permission of the Director of Radio Research c the Department of Scientific and Industrial Research.

(9) REFERENCES

(1) PRESSEY, B. G., ASHWELL, G. E., and FOWLER, C. S.: 'Th Measurement of the Phase Velocity of Ground-Wav Propagation over a Land Path', *Proceedings I.E.E.*, Pap No. 1438 R, March, 1953 (**100**, Part III, p. 73).

(2) ASHWELL, G. E., and FOWLER, C. S.: 'Equipment for th Measurement of Phase Change with Distance of Lo Frequency Ground Waves', *Wireless Engineer* (to b published).

(3) PRESSEY, B. G., and ASHWELL, G. E.: 'The Deviation of Lo Frequency Ground Waves at a Coast-line', *Proceedin I.E.E.*, Paper No. 2083 R (see next page).

(4) CLEMMOW, P. C.: 'Radio Propagation over a Flat Ear across a Boundary Separating Two Different Media *Philosophical Transactions*, A, 1953, **246**, p. 1.

21.396.11

Paper No. 2083 R
July 1956

THE DEVIATION OF LOW-FREQUENCY GROUND WAVES AT A COAST-LINE

By B. G. PRESSEY, M.Sc.(Eng.), Ph.D., Member, and G. E. ASHWELL, B.Sc.

(*The paper was first received 7th December, 1955, and in revised form 17th February, 1956.*)

SUMMARY

After consideration of the methods which have been suggested for omputing the deviation of ground waves at a coast-line, the pheno- nenon is re-examined in the light of recent experimental and theoretical ork on the phase disturbances at such a boundary. It is shown that ie deviation may be calculated from the rate of change of phase with istance along the path of propagation. The changes in this rate hich occur at the boundary give rise to a considerable increase in the agnitude of the deviation as the receiving point is brought within a w wavelengths of that boundary. This increase near the coast seems o provide an explanation of the unexpectedly large deviations reviously observed at medium frequencies.

A series of simultaneous measurements of the phase change and the eviation at 127·5 kc/s along a number of paths crossing the south oast of England are described. Although general agreement between he measured deviations and those derived from the phase curves was btained on some paths, there were appreciable discrepancies on thers. These discrepancies are attributed to the irregularities in the hase surface which were evident over the area and which the method f derivation did not take into account.

(1) INTRODUCTION

Since the first publication by Eckersley[1] in 1920 of experimental evidence of coastal deviations, or coastal refraction as it s commonly called, several theoretical explanations of the henomenon have been put forward. Eckersley's own explanation was in terms of a Zenneck wave. From Zenneck's analysis he derived an expression for the ratio of the velocity of the wave ver sea to that over land in terms of the constants of the medium. Using an exceptionally low value of ground conductivity he btained a value for the ratio which was in agreement with that alculated from the directional measurements using Snell's law. However, the Zenneck expression for the velocity ratio was ncorrectly quoted and the value actually obtained was equal to ts reciprocal. As Smith-Rose[2] has pointed out, Zenneck's analysis leads to a velocity ratio which is less than unity, i.e. the velocity over sea is less than that over land. The explanation in terms of a Zenneck wave thus gives a deviation of the wave in the opposite direction to that observed experimentally. Moreover, f the deviation is calculated from Zenneck's formula using more appropriate values for the ground constants, it is found that the deviation is not only of the opposite sign but also considerably smaller than that obtained experimentally by Eckersley, and also by Smith-Rose.[3] Since it has now been established that the wave system set up by a small radiator on the earth is not a Zenneck wave, an explanation in such terms cannot be accepted.

Grunberg appears to be the first to have suggested a theory o explain coastal refraction. In his 1942 paper[4] the case of a plane wave was discussed, but a simple formula was given only for the case when the observer is at a large distance from the coast. Feinberg[5] generalized this work and derived a simple formula for the deviation which is valid when the transmitter and receiver are at finite distances from the coast. However, the use of the formula is restricted, since the distances must not be a small fraction of a wavelength nor must the corresponding numerical distances be large.

After reconsidering the problem in 1947, Eckersley[6] suggested that the contour of the land at the coast might be an important factor in the explanation of coastal deviation. He showed that, if the land rose from the sea in a hill of the order of 2 km radius, the decrease in velocity of waves passing from sea to land, due to the diffraction over the hill as well as to the lower conductivity, would be sufficient to account for the deviation which he had observed. But, apart from the question of the existence of such a hill, Eckersley himself considered that the theory was not really valid under the conditions of these particular measurements.

As a result of their examination of the relation between the Sommerfeld theory of propagation over a flat earth and the theory of diffraction at a straight edge, Booker and Clemmow[7] come to the conclusion that so-called coastal refraction is really a diffraction effect confined mainly to a region within $\lambda/2\pi$ of the coast. This conclusion suggests that 'coastal refraction' is a misnomer and a more general term, such as 'coastal deviation' (which is used in the present paper) would be preferable.

Except for the investigations of Alpert and Ghorozhankin[8] in 1945, little experimental work appears to have been done on the problem since 1928. The existence of coastal deviations was, however, generally accepted by then and the variation with angle of incidence and with frequency (over the medium-frequency band) were known. Alpert and Ghorozhankin introduced a new approach to the problem by exploring the phase distribution over sea of the radiation from a transmitter situated on the shore. They plotted on a map the equiphase lines and found that the distortions of these lines were in good agreement with the errors in bearing of a transmitter at the phase-measuring station as recorded by a direction-finder sited near the main transmitter on the shore. The results are, however, in complete disagreement, as regards both magnitude and sign, with the earlier observations of Eckersley and Smith-Rose. It is to be noted, however, that the shore transmitter was within 100 m of the water-line and only 2 m above it, and that at the frequency used—about 500 kc/s—the penetration depth (at which the signal is 37% of its value at the surface) for average conductivity soil is about 10 m. Thus it is very probable that the effective coast-line was on the landward side of the transmitter site and that no coastal-deviation phenomena as such were present.

This new line of attack, however, is a promising one and is similar to that adopted by the authors. The series of phase measurements on low-frequency waves[9,10] which have been carried out by them in recent years provide a quantity of data on the phase disturbances at boundaries. It is the purpose of the paper to consider the phenomenon of coastal deviation in the light of these data and recent theoretical work, first in a general manner, then with particular reference to previous measurements, and lastly in comparison with some directional observations which were made simultaneously with some of the above-mentioned phase measurements.

Written contributions on papers published without being read at meetings are nvited for consideration with a view to publication.
The paper is an official communication from the Radio Research Station, Department of Scientific and Industrial Research.

(2) DETERMINATION OF THE DEVIATION FROM THE PHASE CHANGE ALONG THE PATH OF PROPAGATION

The derivation of the deviation using Snell's law assumes that the wave velocity, and hence the rate of change of phase with distance, changes abruptly at the boundary. If the velocities appropriate to propagation over the two media are calculated from Sommerfeld's analysis on the assumption that the media are infinite in extent, the deviation obtained is in the same direction as that observed but of appreciably smaller magnitude. This discrepancy may be explained by the fact that the velocities used are only applicable to points far from the boundary, whereas most recorded observations were made within a few wavelengths of it. From recent theoretical and experimental work it is known that there are considerable phase disturbances near a boundary, and these must be taken into account in the determination of the deviations.

(2.1) Phase Change along the Path

Recent theoretical work by Clemmow[11] and the application by the present authors[9] of Millington's method[12] of field-strength determination over a mixed path to the determination of the phase also has shown that there are appreciable phase disturbances on the far side of a boundary which extend to several wavelengths beyond it. This has been confirmed by extensive measurements at low frequencies over all-land and over land–sea paths.[9,10] The magnitudes of these disturbances vary with frequency and with the ground constants on either side of the boundary. For a boundary between land of medium conductivity and sea, curve (*a*) in Fig. 1 is typical of the general

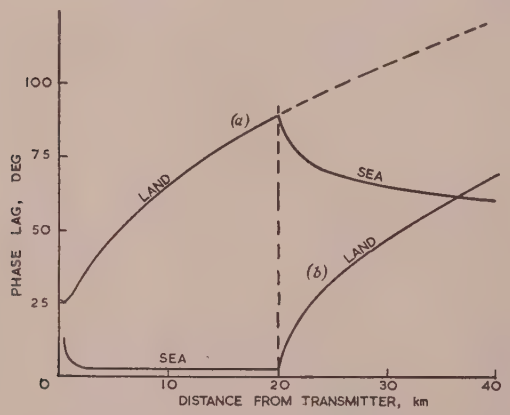


Fig. 1.—Phase change with distance of a wave propagated across a coast-line.

Frequency, 500 kc/s.
Land conductivity, 3×10^{-14} e.m.u.

change of phase with distance of a wave propagated across the boundary from the landward side. In order to emphasize the effects of the ground the phase change has been plotted relative to that which would occur over ground of perfect conductivity; the upward slope of the land portion of the curve indicates an increasing phase lag with distance and hence a velocity lower than that over perfect ground. The main boundary effect shown by this type of curve is the sudden recovery of phase immediately after the wave has crossed the boundary, indicating a sudden change of the instantaneous velocity from a value lower than that over perfect ground to one appreciably higher. There is then a slow decrease of velocity to a value characteristic of propagation over sea. The type of phase change which takes place along a path crossing the same boundary in the opposite direction (i.e. sea to land) is shown by curve (*b*) in Fig. 1; here there is a sudden decrease of velocity on crossing the boundary

(2.2) Method of Deriving the Deviation

The deviation is obtained from these phase curves in the following manner. Consider a wave from a transmitter at T on land propagated across a coast-line AB at an angle α to the normal [Fig. 2(*a*)]. The total phase change with distance may be expressed as

$$\frac{\omega r}{c} + \phi(r)$$

where ω is the angular frequency, r is the distance from the transmitter, c is the velocity in air and $\phi(r)$ is the additional

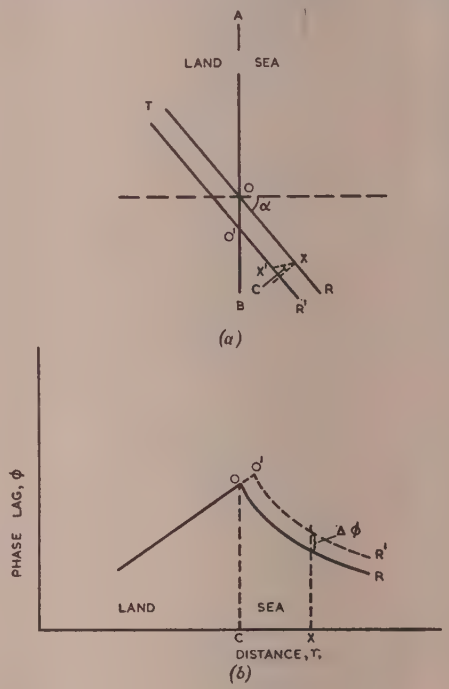


Fig. 2.—Method of deriving the deviation.

(*a*) Propagation paths.
(*b*) Phase changes along paths.

phase lag due to the presence of the ground and is represented by the unbroken curve in Fig. 2(*b*). If now we consider a second path, TR′, parallel to the first, the corresponding phase curve will be as shown by the broken line in Fig. 2(*b*). The distance between the paths is very small, so that the slope of the land portion of the phase curve is the same at O′ as at O and the sea portion O′R′ has the same shape as the portion OR. At a point X the difference between the phase changes along the two paths is $\Delta\phi$, as shown. If we now determine a point X′ on the second path at which the phase is equal to that at X on the first, we may derive an expression for the angular deviation, ϵ, of the phase front from the normal to the paths. The expression obtained is

$$\tan \epsilon = \frac{\lambda}{2\pi} \tan \alpha \left[\left(\frac{d\phi}{dr} \right)_c - \left(\frac{d\phi}{dr} \right)_x \right]$$

where $(d\phi/dr)_c$ and $(d\phi/dr)_x$ are the slopes of the phase curve at the boundary on the landward side and at the point X respectively, and α and ϵ are measured in a clockwise direction from their respective normals.

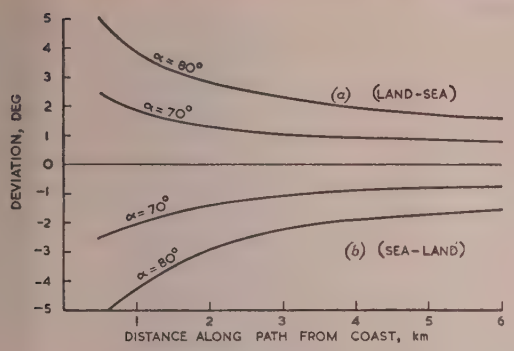


Fig. 3.—Deviation derived from phase change.

The deviations obtained by the above method for specific angles of incidence using the phase curves in Fig. 1 are shown in Fig. 3 plotted on a base of distance along the path from the coast. The main feature of these curves is the rapid increase in the deviation as the coast-line is approached. This is due to the sudden decrease or increase of the phase slope immediately beyond the boundary. If this boundary effect is ignored and the deviation is calculated using the slope which the phase curve approaches at many wavelengths from the boundary, then, as the deviation curve shows, a considerably smaller value will be obtained. It is to be noted that for distances up to 6 km (10λ) from the boundary the deviation is appreciably greater than the asymptotic value so that the effect cannot be regarded as a local one. It is probable that many medium-frequency direction-finding stations situated near the sea lie well within this distance of the coast.

It is of interest to note that, for propagation over a plane earth and within a certain range of frequencies and land conductivities, the above expression for the deviation can be shown to be equivalent to that given by Feinberg[5] for a transmitter at a great distance from the coast.

(2.3) Effect of the Phase Disturbances within a Wavelength of the Boundary

Detailed measurements at 127·5 kc/s of the phase changes across a boundary have shown that within a wavelength of the boundary and on both sides of it there are large phase changes which give rise to high phase slopes. The nature of the changes are shown by the phase curve in Fig. 4, which is typical for

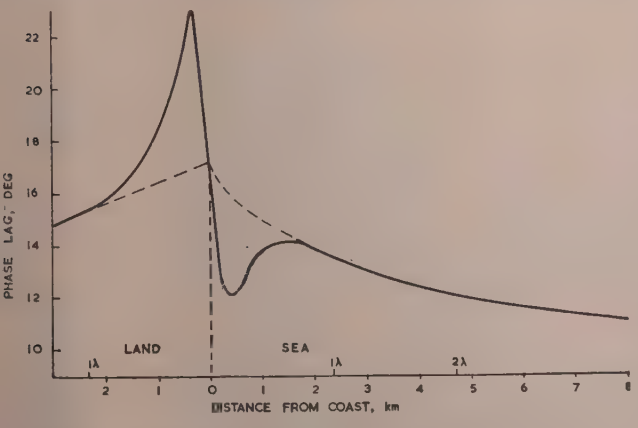


Fig. 4.—Phase change within a wavelength of a boundary.
Frequency, 127·5 kc/s.

transmission from ground of low conductivity to ground of high conductivity along the normal to the boundary. Such changes are not revealed by theoretical or empirical methods of determining the phase change, but it is not unreasonable to assume that changes of a similar nature are also present on higher frequencies. Consideration of the magnitudes of the phase slopes involved leads to the conclusion that the deviations will be correspondingly large, will occur before the boundary as well as beyond it and will suffer some abrupt changes in sign. The deviations before the boundary will presumably correspond to the errors generally experienced on direction-finders situated in the vicinity of a reflecting body.

In the derivation of the deviation curves in Section 2.2 and in Section 2.4 these local phase disturbances have been neglected, because it was assumed that they would not be effective beyond a wavelength from the boundary. It is possible, however, that at small inclinations of the path to the coast this local effect may extend further from the boundary than the half-wavelength shown in Fig. 4, but it is expected that its amplitude will be less and it seems unlikely that it will seriously affect the deviation curves beyond a wavelength from the coast.

(2.4) Application to Existing Data

There appear to be available very few experimental data on coastal deviation which are of a sufficiently precise nature to provide a check on the above theory. The results of Eckersley[1] and those of Smith-Rose[3] are, however, worth examining.

Eckersley's measurements were made on a direction-finding station in Cyprus. The station was situated within a mile of the coast, which ran nearly due north and south in the neighbourhood. He showed the results of bearings taken over a period of five months on transmitters operating in the frequency range 270–375 kc/s. A plot of the differences between the observed and the true bearings against the tangent of the angle of incidence to the coast showed a linear relationship. Unfortunately for our purposes the precise distance of the station from the coast is not given, but in the following analysis a figure of 1 km is assumed. In his analysis Eckersley used a figure of 1×10^{-15} e.m.u. for the conductivity of the ground, but he concludes that it was probably too small; values of $0{\cdot}5 \times 10^{-14}$ and 1×10^{-14} e.m.u. are used here.

The theoretical phase-change curves for propagation at 300 kc/s over a sea–land boundary using the above two values of land conductivity have been drawn, and from them the deviations at a distance of 1 km from the coast and for various angles of incidence have been derived. The coast-line was assumed to be straight, and due allowance was made for the increase in the length of path over land with increase of angle of incidence. The deviations so obtained are shown in Fig. 5, together with Eckersley's results, and it will be noted that there is good agreement between them. For comparison there is also plotted the deviation at a large distance from the coast, where the wave velocity has assumed the value appropriate to land of infinite extent; these latter values are less than one-third of the measured ones.

Smith-Rose's observations were made on a direction-finder at Orfordness, using ship-borne and land-based transmitters operating mainly in the frequency range 500–670 kc/s. For angles of incidence greater than 70° he obtained deviations between 3° and 4°. The phase-change curve appropriate to these conditions would be very similar to curve (*b*) in Fig. 1. Although it is difficult to fix precisely the position of the effective coast-line, it was probably within 1 km of the direction-finder. It will be seen from Fig. 3 that at this distance the derived deviation is between 2° and 4° for angles of incidence between 70° and 80°.

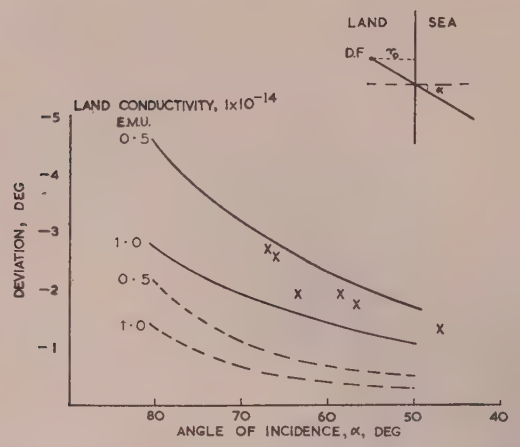


Fig. 5.—Calculated deviations for conditions corresponding to Eckersley's measurements in Cyprus.

× × × × Eckersley's measurements.
——— $r_0 = 1$ km } Calculated.
— — — $r_0 > 20$ km }

The deviation at a great distance from the coast would be less than 1°.

From the above examples it would appear that when the phase changes in the neighbourhood of the coast are taken into account an explanation of the relatively large observed deviations can be obtained.

(3) DEVIATION AND PHASE MEASUREMENTS AT 127·5 KC/S

During a series of measurements of the phase change along paths crossing a coast-line[10] simultaneous directional observations were made at the measuring points at sea. The results provide a means of comparing directly the deviations derived from the phase-change curves with those obtained experimentally.

(3.1) Method of Measurement

The deviation measurements were made with a rotating screened-loop direction-finder incorporating a special optical arrangement for direct observation of the deviation. This consisted of a telescope fixed on the loop mounting with its axis parallel to the axis of the loop. A graticule calibrated in degrees of rotation of the loop from −10 to +10 was mounted in the telescope eyepiece. The observations were made by swinging the loop through about ±3° and observing, at minimum output signal, the reading on the graticule at which a known fixed object on the coast appeared. When the boat and the fixed object were in line with the transmitter this reading gave the deviation directly, but when such an alignment did not exist the necessary corrections could be readily made, since the positions of the boat and the landmark were accurately known.

This method of observing the bearing eliminates the difficulties usually encountered when the bearings are taken with respect to the axis of the boat and corrected for the boat's heading. It was found that, even when the heading of the boat was swinging by as much as ±10° while the bearing was being taken, it was possible to obtain an accuracy of better than ½°.

The boat was constructed mainly of wood and the direction-finder was mounted at the stern so as to be clear of aerials and guy wires which could cause errors. No calibration of the installation was made, but this was unnecessary as all observations were made on a fixed azimuth, i.e. when the boat was heading away from the transmitter. Under these conditions the errors due to the boat or its superstructure were negligible.

When comparing the bearing observations obtained on a direction-finder of this type with values derived from phase data, it must be remembered that the loop position for minimum output signal may not necessarily indicate the direction of the equi-phase surface of the wave at the point of observation. In the first place, the output of the loop is determined by the amplitude gradient as well as by the phase gradient of the field, and if the equi-amplitude contour does not coincide with the equi-phase contour a true indication of the direction of the latter will not be obtained. The possibility of errors being caused in this way under the condition appertaining to these particular measurements has been examined, and it is found that, even close to the boundary where the angle between the equi-amplitude and equi-phase contours is likely to be greatest, the errors will be negligible. Secondly, the field near a boundary may be regarded as the resultant of the direct field and a secondary field originating at the boundary. Consideration of the form of this secondary field suggests that the relation between the electric and magnetic components of the resultant field may be abnormal in the immediate vicinity of the boundary and an erroneous indication of the direction of the equi-phase surface by a loop would not be unexpected. It appears likely, however, that serious errors

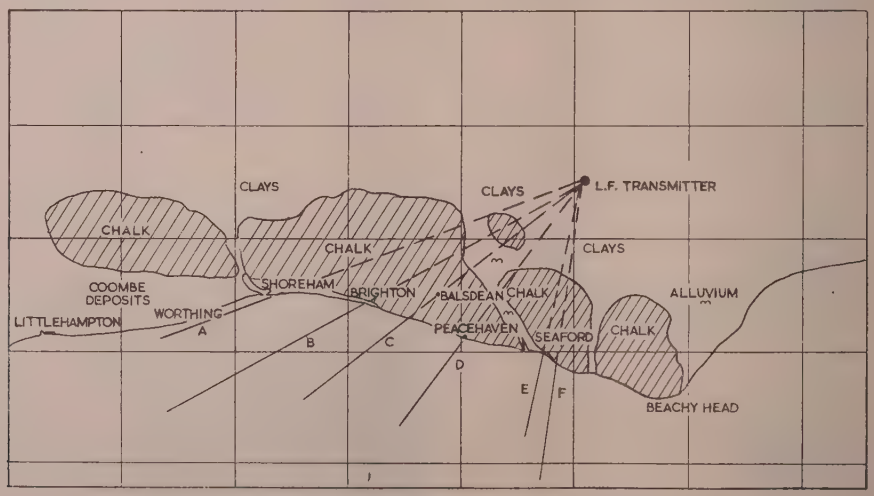


Fig. 6.—Geological map of area showing paths of measurements.

will be confined to within a small fraction of a wavelength of the boundary.

(3.2) Results

Fig. 6 shows a geological sketch map of the area between Beachy Head and Worthing on which are marked the positions of the transmitter and the six radial paths over which reliable directional observations were obtained; the measured deviation on all paths are shown in Fig. 7. The readings were taken at approximately 100 m intervals, and on most paths more than one run was made, the curve shown being the mean for the runs. The observations taken at the same distance along the path on different runs did not differ by more than $\frac{1}{2}°$, and part of this difference may have been due to differences in the lateral positions of the observational points rather than to error of measurement. The smooth curves (shown dotted), which have been added to emphasize the general trends, approach a constant value of between $-0{\cdot}5°$ and $-1{\cdot}5°$ at sea, whereas the expected value is between $+0{\cdot}5°$ and $-0{\cdot}5°$ according to the angle of incidence. This difference is assumed to be due to an error in the alignment of the loop and telescope, which, owing to the conditions of observation, could not be checked precisely and may have differed from path to path.

(3.3) Discussion of Results

Except in the case of path B there is no clear indication of a variation in the deviation such as is shown by the curves in Fig. 3. However, the measured phase-changes did not follow closely the theoretical curve, so that it is essential to examine separately the deviations along each path in relation to their corresponding measured phase curves.*

(3.3.1) Detailed.

Path A (8).—The deviation shows a variation of up to $\frac{3}{4}°$ about a mean value of $-1°$. The measured phase change over this region was constant to $\pm 1\frac{1}{2}°$, so that little change in the deviation would be expected. The absence of a phase-recovery effect at the coast was probably due to the effective conductivity boundary being at the junction of the chalk and Coombe deposits, 7 km inland, rather than at the actual coast-line.

Path B (6).—There is a marked increase in the deviation on approaching the coast. The measured phase change for this path was in close agreement with the theoretical one and the deviation computed from it is shown in the Figure. Although the general level of the measured curve is less than the computed one, for the reason already given, the total change in the deviation with distance is comparable, being about $1{\cdot}2°$ measured and $0{\cdot}9°$ computed. The theoretical value at a large distance from the coast is $0{\cdot}25°$.

Path C (5).—The very large changes in the deviation near the coast were amply confirmed by the close agreement between the results of the two runs made on this path. They are discussed in more detail below.

Path D (4).—Since this path lies nearly perpendicular to the coast-line, no appreciable deviation, or change of deviation with distance, would be expected. Although the general level of the deviation remains constant with distance there are marked variations of up to $\frac{3}{4}°$ about it. The measured phase curve for this path also shows marked variations about the theoretical one.

Path E (3).—The deviations show a negative trend within 3 km of the coast which is greater than that derived from the phase curves. In calculating this deviation the effective boundary has been taken as that situated between the chalk and the alluvium, which is at 45° to the path. Although the measured phase-change showed a very rapid increase (considerably greater than the theoretical) as the coast was approached, it was not sufficient to account for all the change in the deviation observed.

Path F (2).—Beyond 2 km from the coast the variations in the deviation are small, showing a slight positive trend as the coast is approached. The phase curve also showed only a small positive trend in this region and gave a derived deviation of less than $\frac{1}{2}°$. Within 2 km of the coast large erratic changes in the deviation were

* These phase curves are given in Reference 10. The path reference number in that paper is given below in brackets.

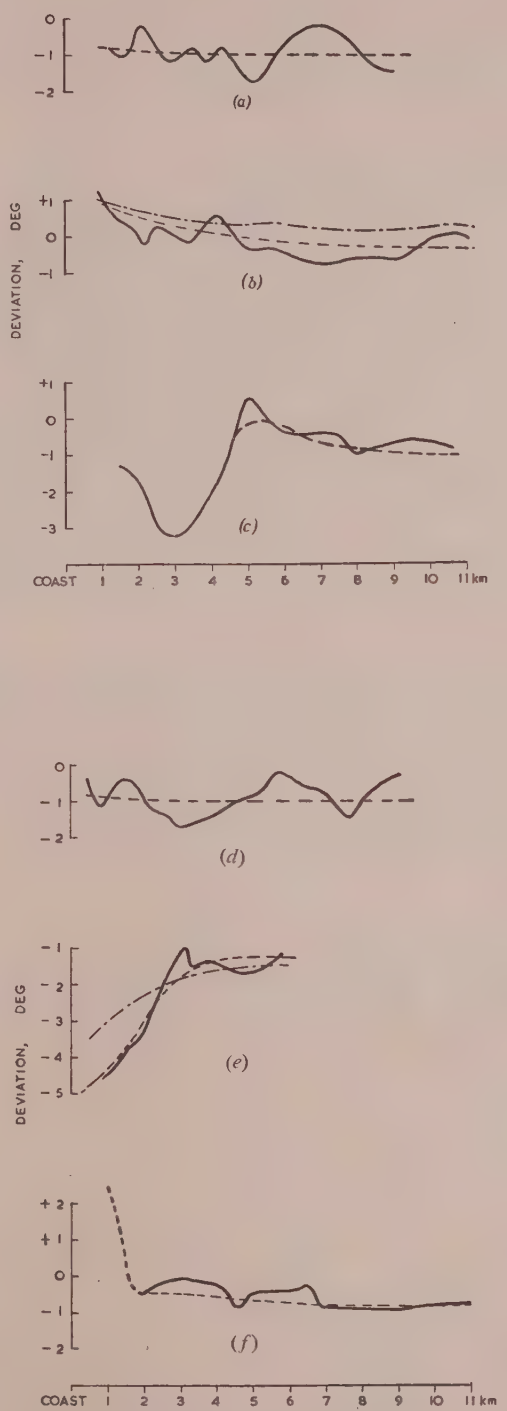


Fig. 7.—Deviations along radial paths.

———— Observed deviation.
— — — — Smoothed observed deviation.
—·—·—·— Deviation derived from radial phase.

(*a*) Path A (Shoreham–Worthing); $\alpha \simeq 60°$.
(*b*) Path B (Brighton); $\alpha = 48°$.
(*c*) Path C (Balsdean); $\alpha = 35°$.
(*d*) Path D (Peacehaven); $\alpha = 15°$.
(*e*) Path E (Seaford West); $\alpha = -45°$.
(*f*) Path F (Seaford East); $\alpha = -30°$.

observed; they were in the positive sense and rose to about 2°, as indicated by the broken line. The phase curve between 0·5 and 2 km from the coast showed an abnormally small change in level, so that the deviation was contrary to the derived value in both sign and magnitude. There was some difficulty in fixing the position of the boat in this area, but it is doubtful whether fix errors could account for the large deviations observed.

(3.3.2) General.

The examination of the observed deviations in relation to the measured phase-changes along the paths has shown that, although the derived deviations may agree with the general trend of the observed changes along some paths, there are large discrepancies along others, particularly those on which there is poor agreement between the measured and the theoretical phase-changes. A closer investigation of these phase anomalies is therefore indicated. In computing the deviations from the phase changes along the path, or radial, one is, in effect, computing the slope of the phase pattern perpendicular to the radial on the assumption that the phase changes along adjacent radials are similar. However, measurements of the phase perpendicular to the paths have shown that this assumption is not valid: the transverse phase slope may differ considerably from that computed from the radial slope. The phase surface is not smooth, in fact, but undulates in an irregular manner. These undulations in the phase surface are undoubtedly the reason for the discrepancies, and better agreement can be obtained only by computing the deviation from the measured transverse phase slope.

There was only one path for which sufficient data were available for determining the transverse phase change. A number of runs had been made perpendicular to this path at various distances along it and also along paths parallel to it. From the results of these measurements it was possible to construct a fairly complete picture of the phase pattern in the region and to deduce the transverse phase slope at many of the points at which directional observations had been made. The deviation was computed from the modified formula

$$\tan \epsilon = \frac{\lambda}{2\pi}\left(\frac{d\phi}{dy}\right)_x$$

where $(d\phi/dy)_x$ is the measured phase slope normal to the path at a distance x from the transmitter. The deviations thus computed are plotted in Fig. 8, together with those computed from the radial phase slope. It will be seen that, while the plot of the latter deviations shows no resemblance to the measured curve, there is good agreement between the measured and computed deviations when the transverse phase slope is used.

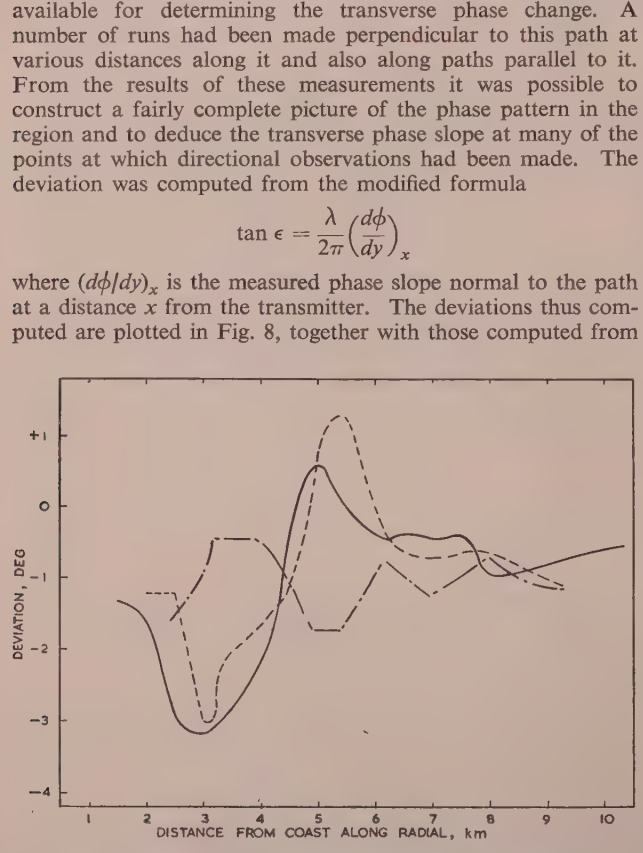


Fig. 8.—Measured and computed deviations along path C.

———— Measured.
— — — — Computed from transverse phase slope.
—·—·— Computed from radial phase slope.

The undulations in the phase surface very probably account in the same way for the variations in the deviation curves on the other paths. This must be so on path D, whose angle of incidence (15°) to the coast is so small that the transverse phase slope corresponding to the radial slope is negligible.

(4) CONCLUSIONS

The deviation of a ground wave near a coast-line has been examined in the light of recent experimental and theoretical work on the phase changes along a path crossing such a geological boundary. It has been shown that the deviation may be calculated from the rate of change of phase, and that the changes in this rate which occur at the boundary give rise to a considerable increase in the magnitude of the deviation as the receiving point is brought within a few wavelengths of the boundary. This increase close to the coast, the existence of which does not seem to have been appreciated hitherto, appears to provide an explanation of the unexpectedly large deviations previously observed on medium frequencies.

From the results of simultaneous measurements of phase and deviation on a frequency of 127·5 kc/s along a number of paths across a coast-line, it is concluded that, where the phase change along the path is in reasonable agreement with the theoretical change, the derived deviations agree with the measured ones. The disagreement between measured and derived deviations which is apparent in the other cases seems to be due to undulations in the phase surface. The presence of these undulations invalidates the relationship between the radial and the transverse phase slopes which is assumed in the calculation of the deviation. If the measured transverse phase slope is used in the calculation of the deviation, close agreement with the measured value is obtained.

One important result of the experimental investigations of phase change with distance and of phase-front deviations has been to reveal the undulating nature of the phase surface near a boundary and to show how large a part it plays in the determination of the deviation. It may be assumed that there will be some degree of undulation in the phase surface near any but the most perfect coast-line, so that deviations derived theoretically on the assumption of a straight boundary with homogeneous ground on either side can be expected to correspond only with the general trend of the measured deviations. If the actual conditions depart too much from the ideal, even this general correspondence may be absent.

The radial paths used in these investigations crossed the coast, or the effective boundary, at angles of incidence of less than 60°. In any further directional measurements it is proposed to use paths which cross the coast at greater angles and on which greater deviations are to be expected as a consequence. A higher-frequency transmitter might also be used, since it would be more convenient for mobile operation and no material decrease in the deviation is expected, provided that the frequency does not exceed about 1 Mc/s.

(5) ACKNOWLEDGMENTS

The authors gratefully acknowledge the valuable assistance given by the Hydrographer of the Navy in the provision of the launch on which the directional measurements were made, and by the officer in command, Lieut. W. D. Stuart, R.N., in the accurate determination of position.

The work described above was carried out as part of the

programme of the Radio Research Board. The paper is published by permission of the Director of Radio Research of the Department of Scientific and Industrial Research.

(6) REFERENCES

(1) ECKERSLEY, T. L.: 'Refraction of Electric Waves', *Radio Review*, 1920, **1**, p. 421.

(2) SMITH-ROSE, R. L.: 'Coastal Errors in Radio Direction-Finding', *Nature*, 1925, **116**, p. 426.

(3) SMITH-ROSE, R. L.: 'A Study of Radio Direction-Finding', R.R.B. Special Report No. 5 (H.M. Stationery Office, 1927).

(4) GRUNBERG, G.: 'Suggestion for a Theory of the Coastal Refraction', *Physical Review*, 1943, **63**, p. 185.

(5) FEINBERG, E.: 'On the Propagation of Radio Waves along an Imperfect Surface—Part IV', *Journal of Physics* (*U.S.S.R.*), 1946, **10**, p. 410.

(6) ECKERSLEY, T. L.: 'Coastal Refraction', *Atti del Congresso Internazionale della Radio*, Rome, September/October, 1947, p. 97.

(7) BOOKER, H. G., and CLEMMOW, P. C.: 'A Relation between the Sommerfeld Theory of Radio Propagation over a Flat Earth and the Theory of Diffraction at a Straight Edge', *Proceedings I.E.E.*, Paper No. 873 R, January, 1950 (**97**, Part III, p. 18).

(8) ALPERT, J. L., and GHOROZHANKIN, B.: 'Experimental Investigation of the Structure of an Electromagnetic Field over the Inhomogeneous Earth's Surface', *Journal of Physics* (*U.S.S.R.*), 1945, **9**, p. 115.

(9) PRESSEY, B. G., ASHWELL, G. E., and FOWLER, C. S.: 'The Measurement of the Phase Velocity of Ground-Wave Propagation at Low Frequencies over a Land Path', *Proceedings I.E.E.*, Paper No. 1438 R, March, 1953 (**100**, Part III, p. 73).

(10) PRESSEY, B. G., ASHWELL, G. E., and FOWLER, C. S.: 'Change of Phase with Distance of a Low-Frequency Ground-Wave propagated across a Coast-Line' (see page 527).

(11) CLEMMOW, P. C.: 'Radio Propagation over a Flat Earth across a Boundary separating Two Different Media', *Philosophical Transactions*, A, 1953, **246**, p. 1.

(12) MILLINGTON, G.: 'Ground-Wave Propagation over an Inhomogeneous Smooth Earth', *Proceedings I.E.E.*, Paper No. 794, January, 1949 (**96**, Part III, p. 53).

621.396.11:621.396.821

Paper No. 2099 R
July 1956

THE PROPAGATION OF A RADIO ATMOSPHERIC

By C. M. SRIVASTAVA, M.Sc.

(*The paper was first received* 19*th March, in revised form* 28*th November*, 1955, *and in final form* 13*th February*, 1956.)

SUMMARY

On the assumption that the space between the earth and the ionosphere acts as a waveguide, the mechanism of propagation of an atmospheric has been considered from the viewpoint of plane-wave reflections. The pulse at the origin has been assumed to be rectangular and of duration 100 microsec. It has been possible to give a physical picture of the mechanism and to explain the oscillatory waveform of distant atmospherics.

(1) INTRODUCTION

Many observers agree that the waveforms of atmospherics can be broadly classified into two types. One type of waveform consists of short irregular impulses which are repeated at intervals that approximately correspond to the repeated reflections of a pulse between the earth and the ionosphere. The other type is oscillatory and is fairly smooth. The former type appears to be well explained by the ray mechanism of multiple reflection. The latter, however, presents certain difficulties when reflection theory is applied for its interpretation. Caton and Pierce[1] have, after an exhaustive study and elaborate discussion, concluded that the interpretation on a single-ray picture is not valid for these waveforms. Budden,[2,3] however, has shown, by mathematical analysis of the propagation of a radio atmospheric between an infinitely conducting earth and an ionosphere of finite conductivity, that the pulse may transform into a smooth oscillation at large distances.

An attempt has been made in the paper to explain the propagation of a radio atmospheric through the space between the earth and the ionosphere on the basis of the optical properties of waveguides. Besides giving a physical picture of the mode of propagation, this method extends the theory of multiple reflection and makes possible the interpretation of smooth oscillatory waveforms of atmospherics on the ray mechanism.

The waveforms of atmospherics depend on two factors:

(*a*) The nature of the pulse generated at the origin.
(*b*) The length and characteristics of the propagation path.

In discussing the frequency characteristic of the propagation path, it is necessary to be precise about the frequency composition of the pulse generated at the origin.

(2) NATURE OF THE SOURCE GENERATING THE PULSE

To investigate the frequency composition of the pulse, we shall consider the nature of the source generating it. Schonland,[4] after an extensive study of lightning discharges by means of Boys's camera, has analysed the progress of a typical lightning discharge as follows:

The discharge from the negative cloud is initiated in the form of a pilot streamer advancing into un-ionized air at the rate of 10^7 cm/sec. After about 50 microsec of travel, a stepped leader starting from the cloud moves towards the tip of the pilot streamer with a speed of 2×10^9 cm/sec, but slows down as it advances and eventually merges into the tip of the pilot streamer. Then after another interval of 50 microsec, if the stroke is long enough, it is caught by a second stepped leader, and so on. As the pilot streamer reaches the ground, one or more positive streamers may be initiated from appropriate conductors which eventually join the pilot streamer, thus linking the earth and the cloud with a highly ionized channel. Over this conducting column a brilliant stroke follows at a speed approaching 10^{10} cm/sec.

(3) FREQUENCY COMPOSITION OF THE PULSE AT THE ORIGIN

The leader stroke is believed to be responsible for predischarges which are frequently observed in the study of the waveforms of atmospherics, while the return stroke gives rise to the main waveform of the atmospheric. At the instant when the return stroke takes place, the highly charged cloud is linked to the earth through an ionized channel for about 100 microsec. In this duration of time the density of ionization falls considerably, so that the channel no longer remains conducting, with the consequence that the discharge stops. This cycle of field changes can be reasonably assumed to give rise to a rectangular pulse of time duration 100 microsec (see Fig. 1). The frequency distribution of such a pulse is readily obtained with the help of Fourier integrals as follows:

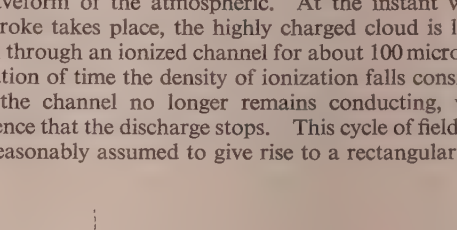


Fig. 1.—Rectangular pulse.

$$F(f) = \int_{-\infty}^{\infty} \exp(-j\omega t) G(t) dt$$

But $G(t) = E$, for $0 \leqslant t \leqslant 10^{-4}$ sec.
$= 0$ for all other values of t.

Therefore $$F(f) = \int_0^{10^{-4}} \exp(-j\omega t) E dt = \frac{E \exp(-j\omega t)}{-j\omega}\Big|_0^{10^{-4}}$$

$$= \frac{E}{\omega}\{\sin(10^{-4}\omega) + j[\cos(10^{-4}\omega) - 1]\}$$

Hence we have the frequency distribution

$$S(\omega) = \frac{2E}{\omega} \sin\left(\frac{10^{-4}\omega}{2}\right) \quad . \quad . \quad . \quad . \quad (1)$$

Fig. 2 gives the curve of $S(\omega)$.

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.

Mr. Srivastava is at the Imperial College of Science and Technology, University of London.

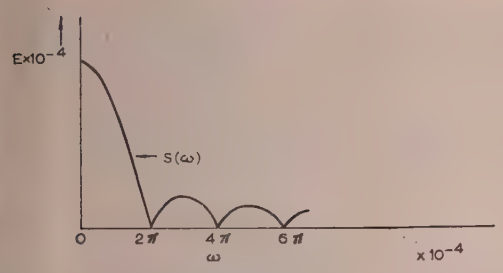


Fig. 2.—Frequency distribution of the components of the rectangular pulse.

(4) CHARACTERISTICS OF THE PROPAGATION PATH

The space between the earth and the ionosphere, through hich the radio atmospheric travels from the initial lightning ash to the receiver, will be treated as a rectangular waveguide ith its two vertical surfaces situated at infinity. In such a aveguide, propagation can take place in several modes which re characterized by the height of the guide alone, the dominant ode being that which has the longest cut-off wavelength. The eld in such a waveguide may be considered as a series of plane aves reflected from wall to wall in the guide in such a manner at they travel in a zigzag path down the guide. This approach ffers a convenient means for visualizing the travelling waves nd endows them with certain optical properties which are of nsiderable help in elucidating the problem of propagation of dio waves.

(4.1) The Optical Properties of a Waveguide consisting of Two Parallel Planes

If the propagation of a wave in the z-direction in a waveguide nsisting of two parallel conducting planes, $x = 0$ and $x = a$ ee Fig. 3), be considered, it can be shown by the method first

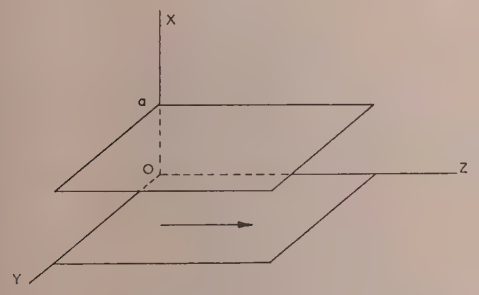


Fig. 3.—Waveguide consisting of two parallel planes.

pounded by Brillouin[5] and later extended by Page and Adams[6] at the E-wave propagated between the two planes is equivalent two plane-polarized waves travelling with the normal velocity f light in the unbounded medium and reflected back and forth tween the planes in a zigzag manner.

In a rectangular waveguide, since all the four walls of the aveguide are effective in imposing boundary conditions for ectromagnetic waves, the propagation takes place in the three-imensional space. However, when the two vertical sides are moved to infinity, the boundary condition for the electro-agnetic waves travelling along the z-direction can be set by e two horizontal planes alone. The propagation can therefore e assumed to take place in a two-dimensional space, independent f the y-co-ordinate. The expression for the field components can therefore be deduced from the expressions for the rectangular waveguide, by rewriting them free of y terms. Thus, for an E_n-wave, the magnetic vector is in the direction of y and the electric field components can be seen[7] to be

$$E_z = \sin\frac{n\pi x}{a}\exp j(\omega t - \beta z)$$

$$E_x = -\frac{i\beta a}{n\pi}\cos\frac{n\pi x}{a}\exp j(\omega t - \beta z) \quad . \quad . \quad (2)$$

where β is the phase constant, and is given by

$$\beta = \left[\left(\frac{\omega}{v}\right)^2 - \left(\frac{n\pi}{a}\right)^2\right]^{1/2}$$

E_x and E_z = Electric field intensities in the x and z-directions, respectively.
a = Height of the waveguide.
ω = Angular frequency of the wave.
n = Order of the mode.
v = Velocity of light in the medium filling the waveguide.

If k, l and m are unit vectors along the axes, the field E can be expressed as

$$E = \left(-k\frac{j\beta a}{n\pi}\cos\frac{n\pi x}{a} + m\sin\frac{n\pi x}{a}\right)\exp j(\omega t - \beta z)$$

$$= -\frac{j}{2}\left[\left(\frac{k\beta a}{n\pi} + m\right)\exp j\left(\omega t - \beta z + \frac{n\pi}{a}x\right)\right.$$

$$\left. + \left(\frac{k\beta a}{n\pi} - m\right)\exp j\left(\omega t - \beta z - \frac{n\pi x}{a}\right)\right]$$

It is evident that it is the sum of two waves travelling in directions given by

$$\cos A = \frac{\pm\frac{n\pi}{a}}{\left[\left(\frac{n\pi}{a}\right)^2 + \beta^2\right]^{1/2}} \quad . \quad . \quad . \quad . \quad (3a)$$

$$\cos B = 0 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (3b)$$

$$\cos C = \frac{\beta}{\left[\left(\frac{n\pi}{a}\right)^2 + \beta^2\right]^{1/2}} \quad . \quad . \quad . \quad . \quad (3c)$$

where A, B and C are the angles the ray makes with the axes x, y and z, respectively.

But since $$\beta^2 = \left(\frac{\omega}{v}\right)^2 - \left(\frac{n\pi}{a}\right)^2$$

we get, from eqn. (3c),

$$\cos C = \frac{v}{\omega}\sqrt{\left[\left(\frac{\omega}{v}\right)^2 - \left(\frac{n\pi}{a}\right)^2\right]} = \sqrt{\left(1 - \frac{n^2\pi^2v^2}{a^2\omega^2}\right)}$$

Now $n\pi v/a = \omega_{0n}$, the critical angular frequency for the nth mode.

Therefore $$\cos C = \sqrt{\left(1 - \frac{\omega_{0n}^2}{\omega^2}\right)} = \sqrt{\left(1 - \frac{f_{0n}^2}{f^2}\right)} \quad . \quad . \quad (4)$$

Eqn. (4) shows that the wave of frequency f can be propagated through the waveguide in the nth mode only at a specific angle C. The actual path is shown in Fig. 4.

Fig. 4.—Path of the rays inside the waveguide consisting of two parallel planes.

Since in such a waveguide the propagation of the electromagnetic wave is possible in a certain number of modes, the received e.m.f. will be found by adding the contributions from all modes.

(4.2) Attenuation Characteristics of the Waveguide formed by the Earth and the Ionosphere

Having determined the propagation characteristics of the waveguide consisting of two infinitely parallel conducting planes, it is necessary to consider how the results already obtained are modified when one of the planes is infinitely conducting while the other is of finite conductivity, as in the case of the earth and the ionosphere.

Since for transmission modes in waveguides consisting of conducting planes the propagation coefficient is purely imaginary, no attenuation of the wave occurs when it travels down the waveguide. However, if one of the surfaces is of finite conductivity, the amplitude of the wave, after it is reflected from the surface, will have $|R|$ times the value it had before it was incident on the surface, while its phase will change by an angle ϕ, where $|R|$ is the modulus and ϕ the phase angle of the complex reflection coefficient $\boldsymbol{R}$ of the surface. Moreover, it may be noted from Fig. 4 that, in order to reach the receiver from the source, the ray of frequency f must make an integral number of reflections at the surface of finite conductivity if the source and the receiver are situated on the surface of infinite conductivity. Hence, if C is the angle that the ray makes with the axes z (Fig. 4), D is the distance of the receiver from the source and a is the height of the guide, we must have

$$\frac{D}{2a}\tan C = N \quad . \quad . \quad . \quad . \quad . \quad . \quad (5)$$

where N is any integer.

But

$$\cos C = \sqrt{\left(1 - \frac{f_{0n}^2}{f^2}\right)} \quad . \quad . \quad . \quad . \quad . \quad (4)$$

Eliminating C from eqns. (4) and (5), we get,

$$f = f_{0n}\sqrt{\left(1 + \frac{D^2}{4a^2N^2}\right)} \quad . \quad . \quad . \quad . \quad . \quad (6)$$

Hence for the ray of frequency f to reach a distance D in a particular mode n, eqn. (6) must be satisfied. For the ray which does not satisfy this condition, the point where the receiver is situated may be considered to lie in the skip distance for the ray.

The transmission characteristics of the path for a certain frequency therefore depends on two factors:

(*a*) The number of modes in which the frequency can be transmitted.

(*b*) The number of reflections which the ray of the frequency under consideration makes at the wall for various modes to reach the receiver.

(5) WAVEFORMS OF ATMOSPHERICS

As has been pointed out earlier, the waveform of the atmospheric is, in fact, that of the e.m.f. received by adding the contributions from all modes. Let us consider the e.m.f. received at the distance D from the origin of the pulse, whose frequency composition at the point of origin is given by eqn. (1).

We have seen that the ray of frequency f in the nth mode must satisfy eqn. (6) if it reaches the receiver after travelling a distance D from the source. From eqns. (1) and (6), the amplitude of the ray of frequency f arriving at the distance D in the nth mode and after making N reflections can be found to be

$$|S(f)|_{n,N} = \frac{k_n E R^N}{\pi f_{0n}\left(1 + \frac{D^2}{4a^2N^2}\right)^{1/2}} \sin\left[10^{-4}\pi f_{0n}\left(1 + \frac{D^2}{4a^2N^2}\right)^{1/2}\right]$$

where k_n is the factor which determines the amplitude of the wave excited by the source in the nth mode. From eqn. (2) it can easily be shown that

$$k_n = \left(1 - \frac{\omega^2 a^2}{v^2 n^2 \pi^2}\right)^{1/2}$$

Further it can be seen that it will be lagging in phase by $\omega\sqrt{(D^2 + 4a^2N^2)}/c$ radians from the phase of the ray at the origin. Therefore, the received e.m.f. in the nth mode will be

$$I_n(t) = \sum_{N=1}^{\infty} \frac{k_n E R^N}{\pi f_{0n}\left(1 + \frac{D^2}{4a^2N^2}\right)^{1/2}} \sin\left[10^{-4}\pi f_{0n}\left(1 + \frac{D^2}{4a^2N^2}\right)^{1/2}\right] \exp\left\{j\omega\left[t - \frac{\sqrt{(D^2 + 4a^2N^2)}}{c}\right]\right\}$$

If we now add the contributions of all the modes, we get the received e.m.f., which should represent the waveform of the atmospheric:

$$I(t) = \sum_{n=0}^{\infty}\sum_{N=1}^{\infty} \frac{k_n E R^N}{\pi f_{0n}\left(1 + \frac{D^2}{4a^2N^2}\right)^{1/2}} \sin\left[10^{-4}\pi f_{0n}\left(1 + \frac{D^2}{4a^2N^2}\right)^{1/2}\right] \exp\left\{j\omega\left[t - \frac{\sqrt{(D^2 + 4a^2N^2)}}{c}\right]\right\} \quad . \quad (8)$$

It is difficult to evaluate eqn. (8) in its present form. We therefore calculate the relative amplitudes of the frequencies that arrive in a particular mode at a considerable distance D (1 000 km) from the source with the help of eqn. (7), and add the contributions of various modes to get the resultant waveform. For the calculation of the amplitudes, $\boldsymbol{R}$ in eqn. (7) should be replaced by its modulus $|R|$. Further, since f_{0n}, the critical frequency, occurs in the denominator, the amplitude will continue decreasing with higher-order modes. The calculation has therefore been made for modes of orders 1 and 2, whose amplitudes are much higher than those of other modes. The height of the ionosphere has been taken to be 70 km. $\boldsymbol{R}$, which is the Fresnel reflection coefficient of the ionosphere for the angle of incidence $(90° - C)$, has been calculated for each frequency. It is given by

$$\boldsymbol{R} = \frac{n_i^2 \sin c - \sqrt{(n_i^2 - \cos^2 c)}}{n_i^2 \sin c + \sqrt{(n_i^2 - \cos^2 c)}} \quad . \quad . \quad . \quad (9$$

where n_i is the refractive index of the ionosphere.

When collisions are taken into account the ionosphere behaves as an absorbing medium of conductivity σ, where

$$\sigma = \frac{N_e e^2 \nu}{m(\omega^2 + \nu^2)}$$

e and m are the electronic charge and mass, ν is the frequency of collision per second, ω is the angular frequency of the wave and

, is the number of electrons per cubic centimetre. As such, has a complex refractive index n_i which is given by

$$n_i^2 = \epsilon = 1 - \frac{4\pi j\sigma}{\omega} \quad . \quad . \quad . \quad . \quad . \quad (10)$$

here ϵ is the complex dielectric constant.
Since in this case $\nu \gg \omega$, we get, from eqn. (10),

$$n_i^2 = 1 - j\frac{4\pi N_e e^2}{m\omega\nu} = 1 - jb \quad . \quad . \quad . \quad . \quad (11)$$

here $$b = \frac{4\pi N_e e^2}{m\omega\nu}$$

Substituting eqn. (11) in eqn. (9) we get

$$= \frac{(1 - jb)\sin c - \sqrt{[(1 - jb) - \cos^2 c]}}{(1 - jb)\sin c + \sqrt{[(1 - jb) - \cos^2 c]}}$$

$$= \frac{(1 + b^2)^{1/2}\varepsilon^{-j\arctan b}\sin c - (\sin^4 c + b^2)^{1/4}\varepsilon^{-j\frac{1}{2}\arctan(b/\sin^2 c)}}{(1 + b^2)^{1/2}\varepsilon^{-j\arctan b}\sin c + (\sin^4 c + b^2)^{1/4}\varepsilon^{-j\frac{1}{2}\arctan(b/\sin^2 c)}}$$

nce in our case $\sin^4 c \ll b^2$

$$R = \frac{\left(\frac{1}{b} + b\right)^{1/2}\varepsilon^{-j(\arctan b - \frac{1}{2}\arctan b/\sin^2 c)}\sin c - 1}{\left(\frac{1}{b} + b\right)^{1/2}\varepsilon^{-j(\arctan b - \frac{1}{2}\arctan b/\sin^2 c)}\sin c + 1}$$

tting

$$\left(\frac{1}{b} + b\right)^{1/2}\cos(\arctan b - \tfrac{1}{2}\arctan b/\sin^2 c)\sin c = A$$

nd $$\left(\frac{1}{b} + b\right)^{1/2}\sin(\arctan b - \tfrac{1}{2}\arctan b/\sin^2 c)\sin c = B$$

e get $$|R| = \frac{\sqrt{(A^4 - 2A^2 + 1 + 2A^2B^2 + 2B^2 + B^4)}}{(A + 1)^2 + B^2} \quad . \quad (12)$$

The values of N_e and ν for the calculation of $|R|$ from eqn. (12) ave been taken from Budden's calculations[8] of the parameters, amely $N_e = 530$ per cubic centimetre, and $\nu = 4{\cdot}2 \times 10^6$ per econd.

Table 1(*a*) gives the calculated frequencies that reach the eceiver after one, two and three reflections at the ionosphere the first-order mode, together with their relative amplitudes. able 1(*b*) gives the same values calculated for the second-order ode.

Table 1(*a*)

Number of reflections	Frequency	Relative amplitudes
	kc/s	
1	15·4	6·52
2	7·9	2·17
3	5·5	0·945

Table 1(*b*)

Number of reflections	Frequency	Relative amplitudes
	kc/s	
1	30·8	1·00
2	15·8	1·64
3	11·1	0·22

It is evident that the frequency of 15·4 kc/s has a much larger amplitude than any other frequency that reaches the receiver. Therefore, if we assume the lightning flash and the receiver to constitute a four-terminal network, the system will have the characteristic of a frequency-selective circuit with a maximum gain at 15·4 kc/s. This agrees fairly well with the observed attenuation curve of the atmospherics.[9]

The problem of finding out the nature of the waveform can now be viewed from the point of view of Fourier-integral analysis. It is equivalent to finding the response of the frequency-selective network to the rectangular pulse whose frequency distribution is shown in Fig. 2. Since the effect of the frequency-selective circuit on the rectangular pulse is to give rise to a 'quasi-sinusoidal' voltage output with a period roughly equal to the cut-off frequency of the network, we can expect that the received waveform is a 'quasi-sinusoidal' oscillation whose quasi-period is approximately 65 microsec at the beginning. The frequency and amplitude of the oscillation will both decrease with time, but it is not possible, at present, to calculate the exact manner in which the decrement would occur.

The period of the 'quasi-sinusoidal' oscillation of a number of waveforms of atmospherics originating from discharges beyond 1 600 km has been observed by Caton and Pierce[1] to increase steadily from 80 to 140 microsec, at the beginning, to about 250 microsec at the end of the waveform. The theory therefore explains on a semi-qualitative basis the experimental observations.

It will now be indicated briefly how, for short distances, a succession of sharp irregular impulses are obtained. Table 2

Table 2

Number of reflections	Order of mode	Frequency
		kc/s
1	1	3·57
	2	7·14
	3	10·71
	4	14·28
2	1	2·57
	2	5·16
	3	7·73
	4	10·31
3	1	2·33
	2	4·66
	3	6·99
	4	9·32

gives the frequencies that arrive at the receiver placed at a distance of 200 km from the source after successive reflections. The height of the ionosphere has been assumed to be 70 km.

It is to be noted that the frequencies that arrive at the receiver after one reflection are octaves of a fundamental frequency of 3·57 kc/s. If we consider a large number of modes, these frequencies will combine to give a sharp impulse.

The frequencies that arrive after two reflections are also octaves of a fundamental frequency, namely 2·57 kc/s, which is very near the fundamental frequency of the first impulse. Hence the second impulse, though more attenuated, has a shape similar to the first impulse.

This method can be extended to explain the other types of waveforms of atmospherics observed by various workers.

(6) ACKNOWLEDGMENT

The author wishes to express his thanks to the Council of Scientific and Industrial Research of India for the provision of a

grant which made possible the work described in the paper. He is indebted to Dr. S. R. Khastgir of the Banaras Hindu University for his many helpful suggestions. His thanks are also due to Dr. Colin Cherry and Dr. J. Brown for having very kindly examined the paper and offered many valuable suggestions.

(7) REFERENCES

(1) Caton, P. F. G., and Pierce, E. T.: 'The Waveforms of Atmospherics', *Philosophical Magazine*, 1952, **43**, p. 393.

(2) Budden, K. G.: 'The Propagation of a Radio-Atmospheric', *ibid.*, 1951, **42**, p. 1.

(3) Budden, K. G.: 'The Propagation of a Radio-Atmospheric II', *ibid.*, 1952, **43**, p. 1179.

(4) Schonland, B. F. J.: 'Progressive Lightning VI', *Proceedings of the Royal Society*, A, 1938, **168**, p. 455.

(5) Brillouin, L.: 'Propagation d'ondes électromagnétiques dans un tuyau', *Revue Générale d'Électricité*, 1936, **40**, p. 227.

(6) Page, L., and Adams, N. I.: 'Electromagnetic Waves in Conducting Tubes', *Physical Review*, 1937, **52**, 647.

(7) Lamont, H. R. L.: 'Wave Guides' (Methuen's Monograph, 1950), p. 68.

(8) Budden, K. G.: 'The Reflection of Very Low Frequency Radio Waves at the Surface of a Sharply Bounded Ionosphere with Superimposed Magnetic Field', *Philosophical Magazine*, 1951, **42**, p. 833.

(9) Gardner, F. F.: 'The Use of Atmospherics to Study the Propagation of Very Long Waves', *ibid.*, 1950, **41**, p. 1259.

1.396.11

Paper No. 2059 R
July 1956

THE PREDICTION OF MAXIMUM USABLE FREQUENCIES FOR RADIOCOMMUNICATION OVER A TRANSEQUATORIAL PATH

By G. McK. ALLCOCK, M.Sc.

(*The paper was first received 27th September, 1955, and in revised form 24th January, 1956.*)

SUMMARY

Times of reception of 15 Mc/s radio waves over a transequatorial path 7 500 km have been recorded throughout the recent period of clining solar activity (1950–54). The analysis of these times has own that predictions of maximum usable frequency made by the ual control-point method were, in general, too high by about 4 Mc/s, d at times by as much as 7 Mc/s or more. This is contrary to the rmal experience for long transmission paths lying within a single misphere.

When a transmission mechanism involving multiple geometrical flections is assumed instead of the forward-scattering mechanism plied by the control-point method, it is found that the path can be nsidered, for the purpose of predicting maximum usable frequencies, consist of three reflections. The discrepancies between prediction d observation, which still remain after a 3-reflection mechanism has en invoked, are attributed mainly to reflections from the sporadic-E gion at the southernmost reflection point, although it is possible that teral deviation of the radio waves is also a contributing factor.

(1) INTRODUCTION

The most efficient use of high frequencies for radiocommunication via the ionosphere depends on the accurate quantitative rediction of the capability of the ionosphere to support transission between any two given points on the earth's surface, for ne complete range of frequencies. In particular, an accurate rediction is required of the maximum usable frequency (m.u.f.), e. the highest frequency that can be used at any given time etween the two terminal points.

The fundamental experimental data from which the m.u.f. is redicted are supplied by over 70 ionospheric observatories cattered throughout the world. These observatories make ourly soundings of the ionospheric layers above them, the xperimental results being obtained in the form of curves of irtual height of reflection of the vertically incident waves plotted gainst frequency (h'f curves). From these curves are derived arameters such as critical frequencies and m.u.f. factors. After nalysis of the global variation of these parameters, diurnally, easonally and throughout the sunspot cycle, predictions can be ade of the m.u.f.'s to be expected for transmission paths of articular interest.

When the length of the path is greater than the maximum length f a single hop via the highest (F2) layer, an assumption must be ade regarding the mode of propagation of the signals. The rocedure adopted by the British[1] and American[2] prediction ervices is to calculate the m.u.f.'s for 4 000 km paths whose eflection points lie on the great circle between the terminals of ne path, at distances of 2 000 km from each terminal. The wer of these two calculated values of m.u.f. is then taken as ne m.u.f. for the circuit. This method is known as the 'control-oint' method, and assumes that transmission of the signals is upported along the path between these control points by some rm of forward scattering.[3] The technique used by the Australian prediction service[4] differs slightly from the above in that the maximum length of a single hop is considered to be 3000 km. Consequently the control points are 1 500 km from the terminals. An alternative point of view is held by the French prediction service,[5] which considers the propagation to be by multiple geometrical hops, with angles preserved on reflection.

Because of the geomagnetic control of the F2-layer,[6] whereby the critical frequency is found to be dependent upon geomagnetic latitude rather than upon geographic latitude, it has been found expedient for prediction purposes to divide the earth's surface into zones. A discussion of prediction errors due to this and other causes has been given by Wilkins and Minnis.[7] The zonal error is usually greatest in the vicinity of a boundary between one zone and the next. The paper describes measurements of times of reception of signals over a long-distance path chosen to lie in the middle of the Pacific I-zone, so that the zonal error should be comparatively small for such a path. A comparison is made between prediction and observation, in relation to the probable mode of propagation along this path, and an explanation of the residual errors is sought in terms of propagation by the sporadic-E layer, and by lateral deviation from the great-circle route.

(2) EXPERIMENTAL DETAILS

The original aim of the experiment was to assess the accuracy of the predictions of m.u.f. supplied by the British, Australian and American prediction services for a path which was considerably longer than a single-hop path, but which was still sufficiently short that the difficulties associated with an antipodal circuit would not be expected to occur. These latter difficulties arise because there are an infinite number of great circles which pass through two antipodal points; any or a range of these great circles may provide the transmission path at a given instant.

The path chosen for this project was from Hawaii to New Zealand. This path is in the middle of the Pacific I-zone and lies almost along a geomagnetic meridian. A convenient signal from Hawaii was provided by the standard-frequency transmission on 15 Mc/s radiated by station WWVH of the U.S. National Bureau of Standards, situated at Kihei, Maui (20·8° N, 156·5° W). The main receiving site was at the Dominion Physical Laboratory, Lower Hutt, New Zealand (41·2° S, 174·9° E), and the great-circle path length was 7 500 km. An earthed vertical aerial approximately one-quarter-wavelength high was used, the received signals being fed into a conventional communication receiver, the a.g.c. voltage from which was recorded via a valve-voltmeter circuit.

The programme from WWVH is interrupted for periods of approximately four minutes immediately after each hour and half-hour. These interruptions served a useful purpose as accurate time-markers, and also allowed an estimate to be made of the degree of interference by signals from WWV, Washington, the other standard-frequency station operating regularly on 15 Mc/s. It was found that on only a very few occasions was the signal strength of WWV sufficient to cause confusion in the interpretation of the records.

Written contributions on papers published without being read at meetings are vited for consideration with a view to publication.

Mr. Allcock is at the Dominion Physical Laboratory, New Zealand.

The period of recording was from August, 1950, to December, 1954 (inclusive)—a total of 53 months, covering continuously the declining portion of the previous cycle of solar activity from a smoothed sunspot number of 70 down to practically zero. Thus a wide variety of ionospheric conditions, arising from a large change in solar activity, was encountered.

Subsidiary observations of signal strength of WWVH, the purposes of which will be discussed later, were carried out at Nandi, Fiji (17·8° S, 177·4° E) and at Waiuku, New Zealand (37·3° S, 174·8° E). The great-circle distances from WWVH of these two additional receiving points were 5 100 and 7 130 km, respectively.

For this experiment the relevant information obtained from the recordings of signal strength of WWVH consisted of skip times, i.e. the times of appearance of the signals in the morning (in-times), and the times of their subsequent disappearance at night (out-times), due to the regular diurnal variation in the density of ionization of the F2-region.

(3) RESULTS

(3.1) Reduction of Data

The appearance of the signals in the morning was almost always abrupt, a large increase in amplitude being observed in the course of a few minutes. The average error in determining in-times was less than ±5 min. On the other hand, as is usually observed at high frequencies, the disappearance of signals at night was generally rather slow, and the average error in determining out-times is considered to be about ±15 min. This contrast in conditions is readily understood when it is realized that the rate of disappearance depends upon the slow decrease in ion density due to decay and dispersive processes in the F2-region, whereas the rate of appearance depends upon the fast increase in ion density owing to the sudden impact of the sun's ionizing radiation upon the region. The same effect is observed in the variation of skip times from day to day; for a typical month the standard deviation of the in-times was about ±0·3 hours, whilst that for the out-times was about ±1·5 hours.

In order to compare results with predictions, monthly median values of skip times were deduced from the daily values. The extent of variation of skip time with change in solar activity and season can be seen by referring to the typical values given in Table 1, where the greatest effect of changing solar activity is shown by the summer out-times.

Table 1

EFFECT OF SOLAR ACTIVITY AND SEASON ON SKIP TIMES OF 15 MC/S SIGNALS OVER A DISTANCE OF 7 500 KM

	Skip time (N.Z.S.T.)		Time difference due to change of solar activity
	Sunspot number = 70	Sunspot number = 0	
Median in-time: Summer	0440	0505	25 min
Winter	0650	0715	25 min
Seasonal difference ..	2 h 10 min	2 h 10 min	
Median out-time: Summer	0245	2200	4¾ h
Winter	1945	1730	2¼ h
Seasonal difference ..	7 h	4½ h	

New Zealand Standard Time (N.Z.S.T.) is 12 hours ahead of Greenwich Mean Time.

(3.2) Failure of Control-Point Method

Month-by-month predictions of the m.u.f. for the path at the corresponding median skip times were made according to the method prescribed by each predicting service. If there were n prediction errors whatsoever, the predicted m.u.f. would alway be equal to 15 Mc/s, the working frequency. The extent t which the British and American predictions are correct in thi case is shown by Figs. 1(*a*) and 1(*b*), where the distributio

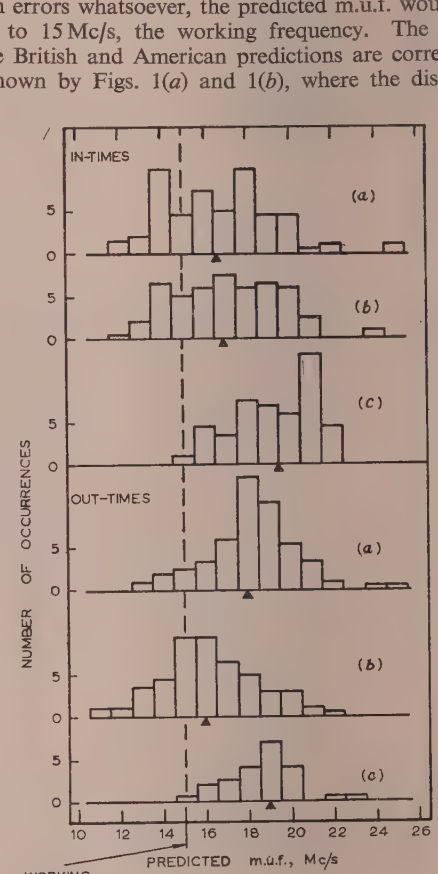


Fig. 1.—Histograms of predicted m.u.f.'s when the working frequenc was 15 Mc/s.

(*a*) British predictions.
(*b*) American predictions.
(*c*) Predictions derived from actual h'f-data using the same control-point metho

The triangles denote median values.

of the predicted values of m.u.f. are shown in histogram form The working frequency of 15 Mc/s is indicated by the dashed line It will be seen that these predictions are on the average abou 2 Mc/s too high, and that there are a number of months in whic the predictions from one or both of the services are high b 5 Mc/s or more. If the predictions of the fundamental quantitie (foF2 and M3 000) had not themselves been generally low, th prediction errors would have been even greater. This is show clearly by the histograms labelled (*c*) in Fig. 1, which are derive from the vertical-incidence h'f-data from ionospheric observa tories in the South Pacific area. Only those months have bee used during which conditions over the southern section of th path exercise control. The remaining months could not b reliably used because of the scarcity of ionospheric observatorie near the nothern section of the path. This is further discusse in the following Section. From Fig. 1(*c*) it is seen that th prediction method, using control points 2 000 km from each en of the path, gave values of m.u.f. which were about 4 Mc/s hig on the average, and occasionally were high by 7 Mc/s or more.

A similar comparison for the Australian predictions is show in Fig. 2(*a*), whilst the corresponding histograms derived fron vertical-incidence h'f-data are shown in Fig. 2(*b*). It is seen tha

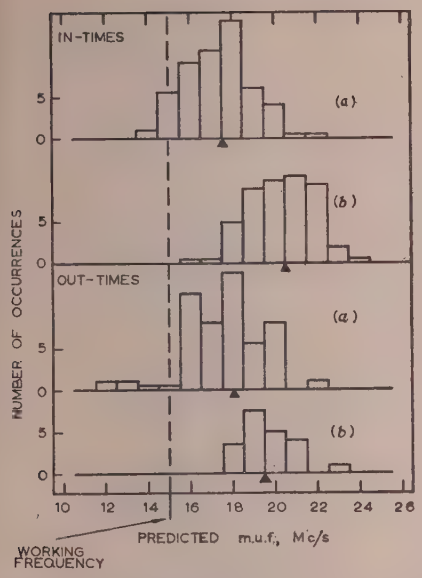


Fig. 2.—Histograms of predicted m.u.f.'s when the working frequency was 15 Mc/s.

(*a*) Australian predictions.
(*b*) Predictions derived from actual h'f-data, using the same control-point method.
The triangles denote median values.

he use of a control-point method, this time with the control oints 1 500 km from each end of the path, produced predictions vhich were always high. It thus appears that the use of control oints is not applicable to this particular path.

It will be noted that the present experience is contrary to that or certain other paths examined by Wilkins and Minnis[7] and by Appleton and Beynon,[8] who found that radiocommunication was ften possible on frequencies exceeding the m.u.f. calculated by he control-point method. A suggested explanation of this pparent disagreement is put forward in Section 3.5.

(3.3) The Multiple-Hop Method

The failure of the control-point method for this path led to an nvestigation of the applicability of a transmission mechanism nvolving multiple geometrical hops. For simplicity the hops vere assumed to be of equal length. This is not as unreasonable n assumption as at first might appear, since a variation of 0 km in the virtual heights of the several reflection points would ave made a difference of only about 200 km in the length of a ingle hop in a practical case. The resulting error in the calulated m.u.f. would have been insignificant compared with the nuch larger errors requiring investigation.

As already mentioned, there was a scarcity of ionospheric bservatories near the northern section of the path. Therefore, efore applying the multiple-hop method to the data, it was ecessary to restrict the analysis to the data for those months uring which it was almost completely certain that conditions t the southernmost reflection point controlled the transmissions. he method of applying this restriction was to examine the elevant h'f-data from the following northern and equatorial bservatories: Panama, Maui, Okinawa, Formosa, Guam, Huancayo and Singapore. Assuming for the time being that he WWVH–Lower Hutt path at the skip time consisted of hree 2 500 km hops, the m.u.f. was calculated for the ionospheric onditions at each of these observatories, the appropriate adjustent in local time being made in each case so that conditions long the nothern section of the path at the actual skip time could be estimated. The data for a given month were used only if the m.u.f. for 2 500 km exceeded 15 Mc/s in all cases. On this criterion, all in-times were found to be subject to control at the southernmost reflection point, but only a limited number of out-times, all near the June solstices, were accepted. A short series of observations at Nandi, Fiji from May to August, 1952, confirmed the conclusion from the above method that during these particular months the southernmost reflection point exercised control.

The reduced number of out-time results, together with all in-time results, were then compared with the ionospheric data from the following South Pacific observatories: Rarotonga, Brisbane, Canberra and Christchurch. The conditions at the appropriate reflection points for two-, three- and four-hop transmission were then estimated, assuming geomagnetic control of the F2-layer, and allowing for the difference in local times at the various reflection points. The m.u.f.'s estimated in this way are plotted in Fig. 3 against the assumed number of hops for the path. For

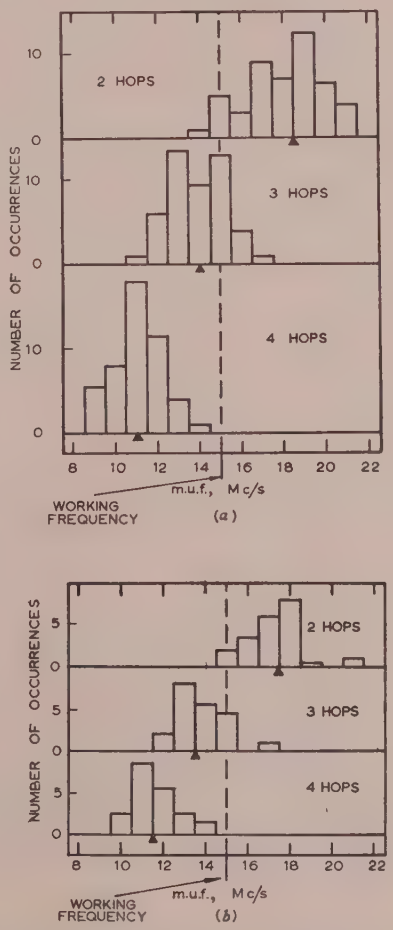


Fig. 3.—Histograms of estimated m.u.f.'s for multiple-hop propagation.

(*a*) Results at in-times.
(*b*) Results at out-times.
The triangles denote median values.

both in-times and out-times the m.u.f.'s calculated on the assumption of a three-hop path are seen to be nearer to the working frequency of 15 Mc/s than the calculated m.u.f.'s assuming either two or four hops. Comparing these results with the control-point results of Figs. 1(*c*) and 2(*b*), it is seen that the assumption

of three identical hops agrees much better with observation than does the control-point method. Furthermore, whereas the calculated m.u.f.'s using the control-point method are too high, the opposite tendency is found for the calculated m.u.f.'s assuming a three-hop path, which are usually somewhat low. This latter state of affairs would be expected if there were other factors contributing to the total transmission of 15 Mc/s signals over the path. Propagation via the sporadic-E region and lateral deviation of the signals from the great-circle path are two such possible additional mechanisms, and are discussed in Sections 3.5 and 3.6.

(3.4) The Maximum Length of a Single Hop

It has been generally assumed that radio waves can be propagated efficiently over a 4 000 km path by means of a single reflection from the F2-layer. On this assumption, it would be expected that two-hop propagation between WWVH and Lower Hutt would be experienced, each hop being 3 750 km in length. With the possible exception of a very few occasions, this does not appear to be the case (Fig. 3). Such two-hop paths were theoretically impossible on some occasions, when the height of the F2-layer was so low that the emission of radio waves tangentially to the surface of the earth would not reach a distance of 3 750 km in a single hop. On the remaining occasions the direction of emission for a two-hop path would be within a few degrees of the horizontal direction, and it would be expected that under these conditions there would be appreciable absorption of energy by the ground in the vicinity of the aerials. Thus it is not surprising that the path appears to consist of three hops at skip times. It is appropriate to recall at this point that the Australian prediction service[9] deduced a maximum practical single-hop length of about 3 000 km from a study of the lowest useful frequencies for the Melbourne–Montreal radiotelegraph circuit. From the present results it seems that the practical limit of a single-hop path lies between 2 500 and 3 750 km—the distances for three-hop and two-hop propagation, respectively. A further series of observations of skip times at Waiuku, New Zealand, 7 130 km from WWVH, from August, 1950, to July, 1951 (inclusive), indicated that this shorter path was also predominantly a three-hop path. Incorporating this latter result, it is inferred that the practical limit of a single-hop path lies between 2 500 and 3 560 km for north–south transmission in latitudes between 20° S and 40° S. This is in general agreement with the Australian experience.

Recent observations by Shearman[11] of ionospheric propagation by means of ground back-scatter have shown that the limiting range for scatter reception is about 3 000 km, which agrees with the above results for one-way transmission. Furthermore, the multiple-hop nature of Shearman's results for scatter signals from very long distances (Figs. 3A and 3B of his paper) does not suggest that forward scattering in the ionosphere plays any significant role in long-distance transmission. This conclusion is also implied by Allan,[12] who found a large degree of coherence in the standard-frequency signals received over the same path as was used in the present experiment.

(3.5) The Influence of the Sporadic-E Region

From Fig. 4 it can be seen that there is a seasonal variation in the accuracy of prediction assuming a three-hop path at in-times. In local winter at the southermost reflection point, the calculated m.u.f.'s can be identified with the working frequency, and it is very likely that a simple three-hop transmission via the F2-layer is the predominant mode of propagation. However, in local summer the calculated m.u.f. is about 2 Mc/s below the working frequency, on the average. This implies the existence during these months of an additional propagation agency which supports

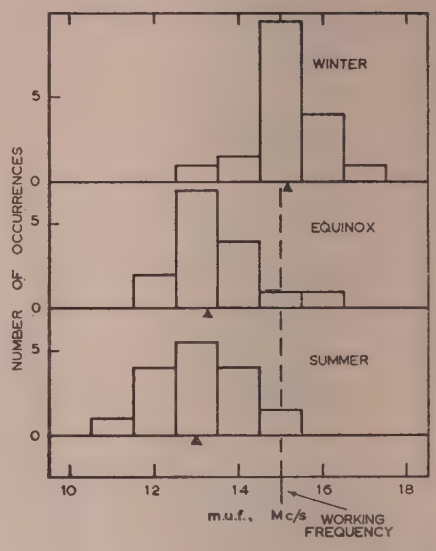


Fig. 4.—Histograms of estimated three-hop m.u.f.'s at in-times showing seasonal variation in prediction accuracy.
The triangles denote median values.

transmission over the southern section of the path for some time before the three-hop m.u.f. rises to 15 Mc/s and regular F2-layer transmission can take place. A similar seasonal variation in discrepancy between calculated and observed m.u.f. has been found by Appleton and Beynon,[8] who have attributed the effect to the sporadic-E region, which occurs mainly during local summer in temperate latitudes.

As there were no ionospheric observatories close to the area of reflection of signals by the sporadic-E region, which would be in the vicinity of the Kermadec Islands, no direct comparison could be made between the occurrence of sporadic-E ionization near the reflecting area and the calculated three-hop m.u.f. However, prior to 1947 an ionospheric observatory was in existence at Kermadec, and a preliminary comparison was made

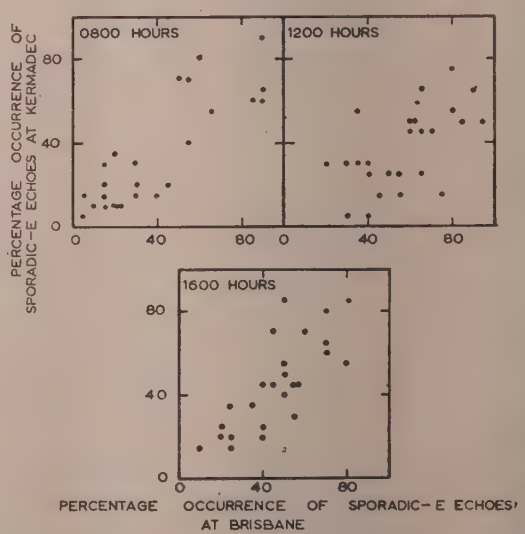


Fig. 5.—Correlation of the occurrence of sporadic-E echoes on 3 Mc/s at Brisbane and Kermadec, at 0800, 1200 and 1600 Local Mean Time (1944–47).

the occurrence of sporadic-E ionization at Kermadec and at risbane for the period 1944–47. These two observatories were n approximately the same parallel of latitude, but were separated y about 2 800 km. When monthly averages are considered, the orrelation between these places is surprisingly good, in spite of ie sporadic nature of the phenomenon. Correlograms of the ercentage of occurrence of sporadic-E ionization detectable at Mc/s are shown in Fig. 5 for local mean times of 0800, 1200 nd 1600 hours. The correlation coefficients for these three times re 0·85, 0·65 and 0·75, respectively. It was therefore con- dered to be reasonably satisfactory to compare measurements f sporadic-E ionization at Brisbane with the calculated three-hop n.u.f. for the WWVH–Lower Hutt path. This comparison is hown in Fig. 6, which clearly indicates that the occurrence of

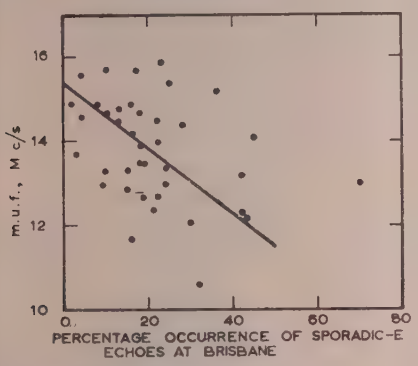


Fig. 6.—Correlation of estimated three-hop m.u.f. at in-times, with occurrence of sporadic-E echoes on 3 Mc/s at Brisbane.

sporadic-E ionization is accompanied by a depression in the calculated m.u.f. It will also be seen that the regression line through the plotted points indicates a calculated m.u.f. of about 15 Mc/s when no sporadic-E ionization is present. This can be interpreted as implying that propagation via the sporadic-E region is the main cause of variation of the WWVH–Lower Hutt path from a simple 3-hop path.

(3.6) Other Possible Causes of Discrepancies

It is likely that lateral deviation of the radio waves from the great-circle path contributes to the discrepancy between predic- tion and observation, signals arriving by a deviated path for some time before the m.u.f. along the great-circle path has risen to the working frequency. As in the case of sporadic-E trans- mission, lateral deviation would tend to raise the actual m.u.f. above the predicted value. A subsidiary experiment to determine the magnitude of this effect was carried out during one week in October, 1952, at Seagrove, New Zealand. The angle of arrival of the signals at in-times was measured, using the rotating interferometer developed by Whale.[10] Small indications were obtained of a swing in the angle of arrival towards the east, i.e. towards a region of higher ionization density produced by the onset of ionizing radiation from the sun. However, this change in bearing angle was found to be less than 3°, which is too small for lateral deviation to be a major cause of discrepancy. It is of interest to record that the vertical angles of arrival measured near the skip times were not inconsistent with the deduction previously made, that the path consisted at that time of three hops.

As has already been noted in Section 3.3, any inequality in the length of hops will introduce a prediction error. A study of the h'f-data from the relevant ionospheric observatories suggests that the error due to this cause is small; nevertheless it is instructive to inquire whether this inequality in hop length is likely to increase or decrease the actual m.u.f. with respect to the predicted m.u.f. for hops of equal length. A radio wave of frequency f_0 approach- ing the m.u.f. is known to be reflected from a higher virtual level in the F2-layer than another wave whose frequency f_1 is consider- ably less than the m.u.f. Thus the great-circle distance travelled by f_0 for a single reflection from the ionosphere will be greater than that travelled by f_1. The effect then consists of lengthening the hop which is due to reflection at that point of the transmission path which determines the m.u.f., and thereby shortening the other hops. Since the results in the paper deal only with cases of control at the southernmost reflection point, this effect pro- duces a shift in the reflection point towards the north, i.e. towards a region of greater ionization density. Therefore the actual m.u.f. will be increased with respect to the predicted m.u.f., as was found to be the case for sporadic-E transmission and lateral deviation.

(4) CONCLUSIONS

The 15 Mc/s standard-frequency transmissions from WWVH, Hawaii, were received and recorded in New Zealand at a distance of 7 500 km from August, 1950, to December, 1954 (inclusive). From the recordings, the skip times have been extracted and compared with predictions of m.u.f. made by the British, Ameri- can and Australian prediction services, and with calculations of m.u.f. made directly from the h'f data from ionospheric observa- tories in the South Pacific area. It has been found that the usual control-point method of making predictions is not applicable to this transequatorial path. The use of such a method led to predicted m.u.f.'s which were, on the average, about 4 Mc/s higher than the observed value. On occasions the predictions were high by at least 7 Mc/s.

The experimental results are consistent with an alternative viewpoint, namely that propagation takes place by three equal- length hops, in which the angles of incidence and reflection are preserved. Predictions made on this assumption are generally 1–2 Mc/s low; this residual discrepancy can be interpreted as being mainly due to the occasional presence of sufficient ioniza- tion in the sporadic-E region to support the transmission of the 15 Mc/s signals over the southern portion of the path from Hawaii to New Zealand.

Some measurements of the lateral deviation in angle of arrival of the 15 Mc/s signals show a swing of up to 3° towards the east at in-times. However, this is not sufficient to account for more than a minor part of the residual discrepancy. The possible presence of hops of unequal length is considered to be still less important.

The above measurements of skip times, combined with a further series of similar measurements at a distance of 7 130 km, indicate that the practical maximum length of a single hop for north–south transmission in southern temperate latitudes lies between 2 500 and 3 560 km, which is not inconsistent with the figure of 3 000 km deduced by the Australian Ionospheric Pre- diction Service.

(5) ACKNOWLEDGMENTS

Observations at Waiuku were made by Mr. O. R. Hull, and those at Nandi, Fiji, by the staff of the New Zealand Civil Avia- tion Administration. Direction-finding facilities at Seagrove were provided by Dr. H. A. Whale, whilst Mr. F. A. McNeill was responsible for most of the routine recording and reduction of skip times at the Dominion Physical Laboratory. The author is indebted to the above for their invaluable assistance, and to the Secretary of the New Zealand Department of Scientific and Industrial Research for permission to publish the paper.

(6) REFERENCES

(1) 'Predictions of Radio Wave Propagation Conditions (RRS Bulletin A)' (Department of Scientific and Industrial Research, London).

(2) 'Basic Radio Propagation Predictions (CRPL Series D)' (National Bureau of Standards, Washington).

(3) U.S. NATIONAL BUREAU OF STANDARDS: 'Ionospheric Radio Propagation' (National Bureau of Standards Circular 462, 1948), p. 74.

(4) 'Ionospheric Predictions (IPS Series W)' (Commonwealth Observatory, Australia).

(5) 'Prévisions pour la Propagation Radioélectrique (SPIM Series A)' (Ministère de la Défense Nationale, France).

(6) APPLETON, E. V.: 'Two Anomalies in the Ionosphere', *Nature*, 1946, **157**, p. 691.

(7) WILKINS, A. F., and MINNIS, C. M.: 'Comparison of Ionospheric Radio Transmission Forecasts with Practical Results', *Proceedings I.E.E.*, Paper No. 1101 R, May 1951 (**98**, Part III, p. 209).

(8) APPLETON, E. V., and BEYNON, W. J. G.: 'Radiocommunication on Frequencies Exceeding Predicted Values', *ibid.*, Paper No. 1461 R, July, 1953 (**100**, Part III, p. 192).

(9) IONOSPHERIC PREDICTION SERVICE: 'Ionospheric Forecasting Errors in Absorption-Limited Circuits' (Commonwealth Observatory, Australia, 1949).

(10) WHALE, H. A.: 'A Rotating Interferometer for the Measurement of the Directions of Arrival of Short Radio Waves', *Proceedings of the Physical Society*, B, 1954, **67**, p. 553.

(11) SHEARMAN, E. D. R.: 'A Study of Ionospheric Propagation by Means of Ground Back-Scatter', *Proceedings I.E.E.*, Paper No. 1914 R, October, 1955 (**103** B, p. 203).

(12) ALLAN, A. H.: 'Variations of Received Frequency of WWVH', *Journal I.E.E.*, 1955, **1** (New Series), p. 650.

621.396.82

Paper No. 2118 R
July 1956

ASYMMETRY IN THE PERFORMANCE OF HIGH-FREQUENCY RADIOTELEGRAPH CIRCUITS

By A. M. HUMBY, Member, and C. M. MINNIS, M.Sc., F.Inst.P., Associate Member.

(*The paper was first received* 10*th January, and in revised form* 5*th March,* 1956.)

SUMMARY

It is sometimes found that the performance of a trans-equatorial diocommunication circuit in the high-frequency band differs conierably, at certain times of day, from that in the opposite direction. iis phenomenon has been referred to as 'circuit asymmetry' and, in e cases considered, it cannot be accounted for in terms of differences the performance of the terminal equipment. Evidence is produced nich tends to show that the asymmetry is sometimes due to the decrease the signal/noise ratio in one direction which can occur when a stant thunderstorm area, lying in the direction of the main beam of e receiving aerial, reaches its diurnal activity maximum.

TERMINOLOGY

Some of the terms used in the paper either have no generally ccepted meanings or are used in a special sense. Such terms e defined here to prevent possible misunderstanding.

Circuit Performance.—The efficiency of a circuit as measured, g., by the number of hours per day or days per month during hich its performance is of commercial grade, i.e. the signals re suitable for printing at high speed.

Percentage Time Commercial.—The percentage time for which ie circuit performance is of commercial grade as defined above.

Asymmetry in Circuit Performance (*Circuit Asymmetry*).—A ifference between the performance of a circuit in the incoming nd outgoing directions.

Seasonal Asymmetry.—The component of the asymmetry which aries regularly with season.

Long-Term Asymmetry.—The slowly varying component of ie asymmetry which remains after subtracting the seasonal omponent.

Lowest Useful High Frequency (*l.u.f.*).—The lowest frequency, etween about 3 and 30 Mc/s, that can be used for a given rvice at a specified time for ionospheric propagation of a radio ave between any two points.

(1) INTRODUCTION

In a recent paper the different factors which influence the erformance of radiotelegraph circuits in the high-frequency .f.) band were reviewed by Humby, Minnis and Hitchcock.[1] n particular, reference was made to several instances in which ystematic differences were found to occur in the performance f a circuit in the outward and inward directions. The term ircuit asymmetry' was introduced to describe this effect and o distinguish it from non-reciprocity, which has a much more recise meaning.[1] A distinction was made between long-term symmetry, which remains almost constant throughout the year, nd seasonal asymmetry, which is the component which varies egularly with season. Both types can occur together, as shown n Fig. 1; in this case the seasonal changes in performance on ie outgoing and incoming circuits are in opposition, but the amplitudes are not great enough, except during winter 1953–54, to outweigh the long-term asymmetry and thereby to cause the circuit performance in the Simonstown to London direction to exceed that in the reverse direction.

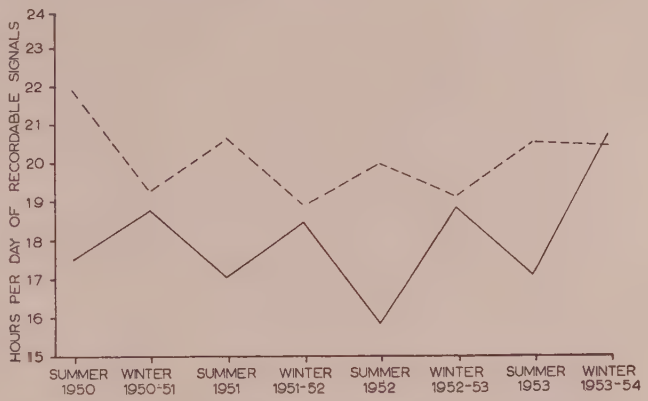


Fig. 1.—Long-term and seasonal asymmetry of the Simonstown (Capetown)–London circuit performance.

---- From London.
—— To London.
Summer: May, June, July.
Winter: November, December, January.

Although Humby, Minnis and Hitchcock mentioned various possible explanations of seasonal asymmetry, they did not find it possible to attribute it definitely to any particular cause. The reason for this failure is that the efficiency of a radio circuit, considered as a channel for passing traffic, is controlled by many different factors, and it is difficult to make an accurate assessment of the relative importance of these in any particular case.

It is well known that the signal/noise ratio in the h.f. band is usually determined by atmospheric noise generated in thunderstorm centres in different parts of the world and often propagated over great distances. The purpose of the paper is to present circuit performance data suggesting that some of the asymmetry effects which have been reported may result from a combination of effects arising from the use of directive receiving aerials and the diurnal and seasonal changes in the sources of atmospheric noise.

(2) EFFECT OF NOISE AT THE RECEIVING TERMINAL

It is evident from Fig. 1 that the seasonal asymmetry on the London–South Africa circuit is such that the performance falls to a minimum during local summer at the receiving terminal. Figs. 2(*a*) and 2(*b*), which refer to the Montreal–Melbourne and London–Harman (Canberra) circuits respectively, show similar minima when the circuit performance is measured in terms of the departure from the mean performance in both directions.

It has been established that there is an annual north–south movement, which follows the sun, of the principal equatorial thunderstorm areas of the world and also that, even in comparatively high latitudes, there is an increase in local thunderstorm

Written contributions on papers published without being read at meetings are vited for consideration with a view to publication.
The paper is an official communication from the Radio Research Station, Depart-ent of Scientific and Industrial Research.
Mr. Humby is now in the Royal Naval Scientific Service.

activity during local summer. In view of these facts it seems reasonable to expect that, in general, there should be an increase in the level of atmospheric noise during local summer at any particular place. As a result, it would be expected that there should be a fall in the performance of a trans-equatorial circuit in the direction which entails reception in local summer. Hence, even if other considerations play a part, it seems likely that atmospheric noise may be partly responsible for seasonal asymmetry on trans-equatorial circuits such as that evident in Figs. 1 and 2.

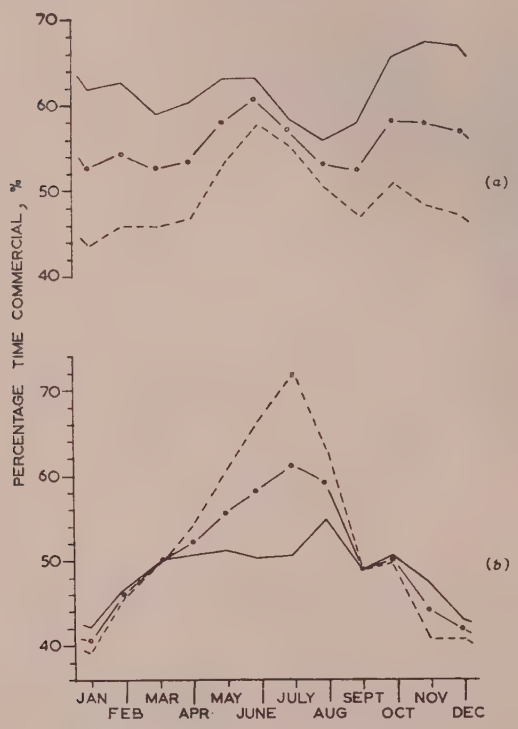


Fig. 2.—Seasonal changes in performance.
(*a*) Montreal–Melbourne circuit (1935–1953 inclusive).
(*b*) London–Harman (Canberra) circuit (1950–1954 inclusive).
—— From Australia.
- - - - To Australia.
—o— Mean of both directions.

Although seasonal asymmetry is frequently encountered on trans-equatorial circuits, it has not been found to occur on circuits in which both terminals are at high latitudes in the same hemisphere. For example, the London–Montreal circuit shows substantially similar seasonal variations in performance in both directions (see Fig. 3); the pronounced semi-annual component of these changes in performance is common to all trans-atlantic circuits and is attributed to the peaks in magnetic activity which occur at the equinoxes (also shown in Fig. 3). Since transatlantic circuits pass near, and run parallel to, the northern auroral zone, they are particularly sensitive to magnetic disturbances.

The diurnal and seasonal changes in the performance of a typical transatlantic circuit during a period of low solar activity are illustrated in Fig. 4. Owing to the comparatively short distance involved, the average performance of the circuit is high and the summer-time increase in noise-level has no apparent effect on the performance. In other seasons the performance remains high during the day, but difficulties are encountered at night because the frequencies required for operation are lower

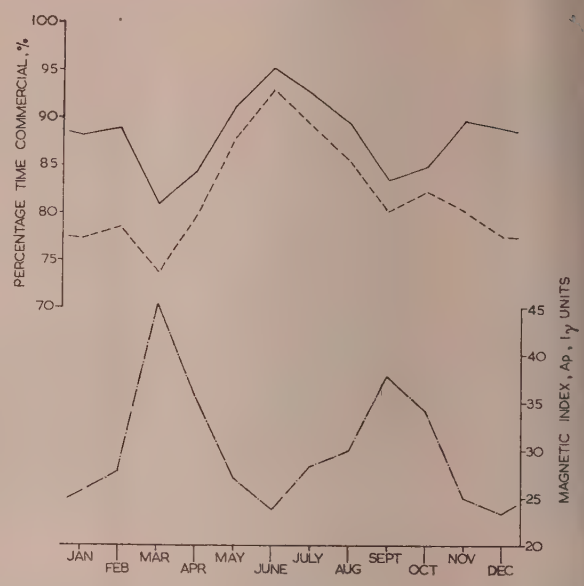


Fig. 3.—Relation between circuit performance and geomagnetic activit on the Montreal–London circuit (1937–47 inclusive).
—— Performance to London.
- - - - Performance from London.
—·— Magnetic index.

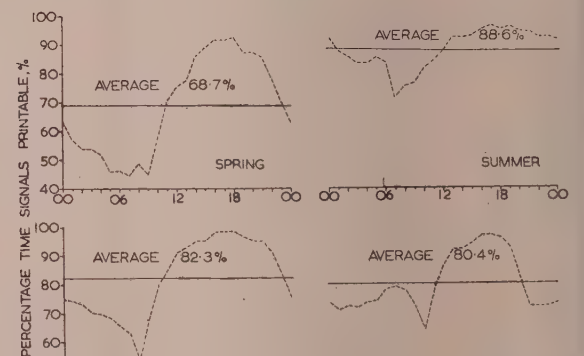


Fig. 4.—Diurnal changes in performance of the London–Halifax circuit (1951–54 inclusive).

than in summer and there is, consequently, less margin for temporary decrease in frequency during disturbed ionospheric o magnetic conditions. Spring, in Figs. 4–8 and 10, include February, March and April, and the other seasons correspond

(3) DIURNAL CHANGES IN CIRCUIT ASYMMETRY

In Figs. 5, 6, 7 and 8, for the circuits from London to Simons town (Capetown), Colombo, Singapore and Harman (Canberra) respectively, the performance is shown as a function of seaso and time of day. On all these circuits asymmetrical performanc is seen to occur at certain times of day, but it is particularl prominent during northern summer, when there is a marked dro in the performance of the circuit incoming to London relativ to that of the outgoing circuit. A striking feature of thi phenomenon is the fact that for each of these four circuits th approximate mean daily time of onset of asymmetry correspond

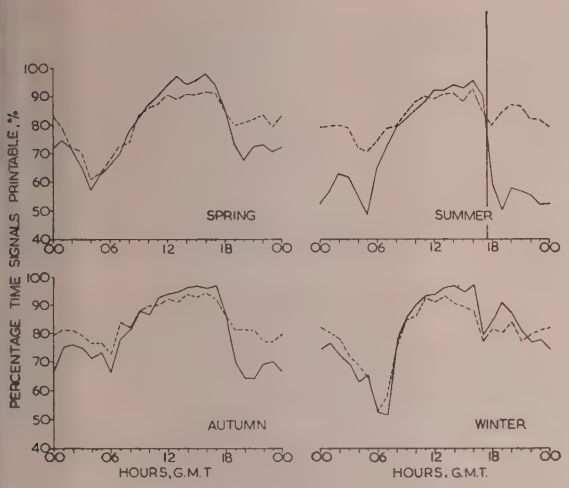


Fig. 5.—Diurnal changes in performance of the Simonstown (Capetown)–London circuit (1951–54 inclusive).

—— To London.
---- From London.
The vertical line indicates time of onset of asymmetry.

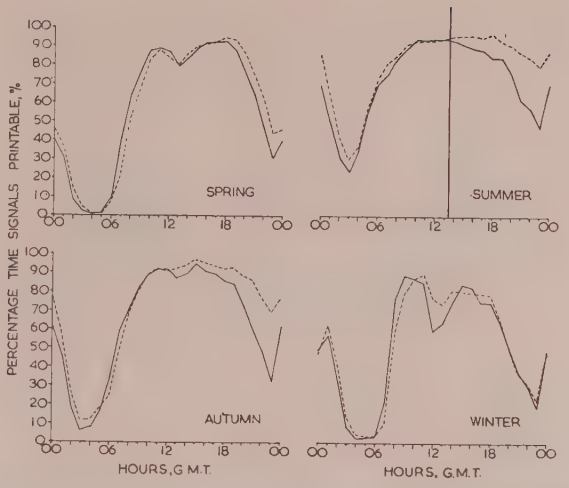


Fig. 7.—Diurnal changes in performance of the Singapore–London circuit (1951–54 inclusive).

—— To London.
---- From London.
The vertical line indicates time of onset of asymmetry.

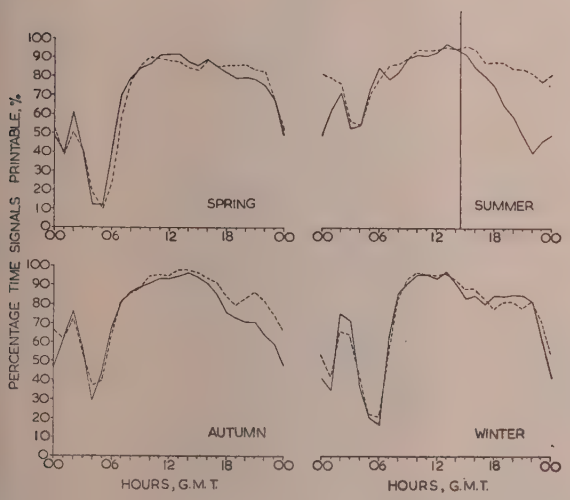


Fig. 6.—Diurnal changes in performance of the Colombo–London circuit (1951–54 inclusive).

—— To London.
---- From London.
The vertical line indicates time of onset of asymmetry.

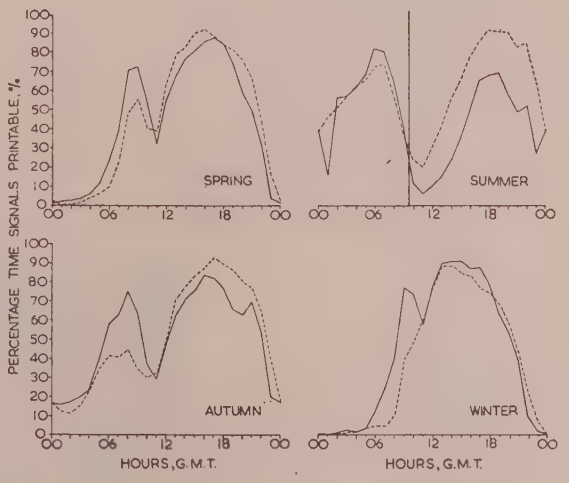


Fig. 8.—Diurnal changes in the performance of the Harman (Canberra)–London circuit (1951–54 inclusive).

—— To London.
---- From London.
The vertical line indicates time of onset of asymmetry

Table 1

ASYMMETRY IN NORTHERN SUMMER FOR CIRCUITS TERMINATING IN LONDON

Distant transmitter			Approximate time of onset of asymmetry		Derived storm location
Location	Longitude	Time ahead of G.M.T.	G.M.T.	L.M.T. at transmitter	
		h min	hours	hours	
Simonstown	20° E	1 20	1730	1850	North/Central Africa
Colombo	80° E	5 20	1430	1950	Europe/N.W. India
Singapore	105° E	7 00	1330	2030	Europe/India
Harman (Canberra)	150° E	10 00	0930	1930	East Indies

fairly closely to a constant local mean time (1850–2030 hours L.M.T.) at the end of the circuit distant from London (see Table 1). A relation of this kind suggests that the onset of asymmetry, which has been indicated in each Figure by a vertical line in the summer record, is associated in some way with the east-to-west movement of the sun.

At this point it is important to remember that directive receiving aerials are invariably used for the circuits under discussion. Although side lobes cannot be completely suppressed, the total energy picked up on them from distributed noise sources is much less than their peak amplitude might suggest. Because of this, the integrated noise level at the terminals of the types of aerial used will usually be determined mainly by the noise sources in or near the direction of the main beam.

If atmospheric noise is, in fact, responsible for the summer afternoon fall in performance shown in Figs. 5, 6, 7 and 8, the geographical locations of the relevant noise sources can be found approximately from a knowledge of the time of onset of circuit asymmetry, provided that two assumptions are made:

(*a*) For a given directive aerial, there is a marked increase in the noise voltage appearing at the aerial terminals when an active noise source moves into the main beam.

(*b*) The activity of all noise sources increases significant between 1600 hours L.M.T. and local sunset.

It has been stated by Brooks[2] that thunderstorm activity ov land reaches its peak at about 1500 hours L.M.T., but, since th matter in question here is the field strength of atmospherics r received in London, it seems likely that the afternoon decrea in ionospheric absorption will cause a delay in the effective tim of peak activity. Some evidence of such a delay is apparent i the data given by Bureau and Bost[3] concerning the reception i Western Europe of atmospherics which originate to the east an south-east.

Given the assumptions mentioned, the intersection of th main beam of the receiving aerial with a zone slightly west o the sunset line can be used to define the general area from whic the atmospheric noise would have to originate if it wer responsible for the onset times given in Table 1. These area which are also included in the Table, cannot be precisely define because they are determined by the intersection of two rathe wide bands. However, they correspond fairly well with area (shown in Fig. 9) which are known from independent evidenc to be the principal centres of high thunderstorm activity durin northern summer.[2]

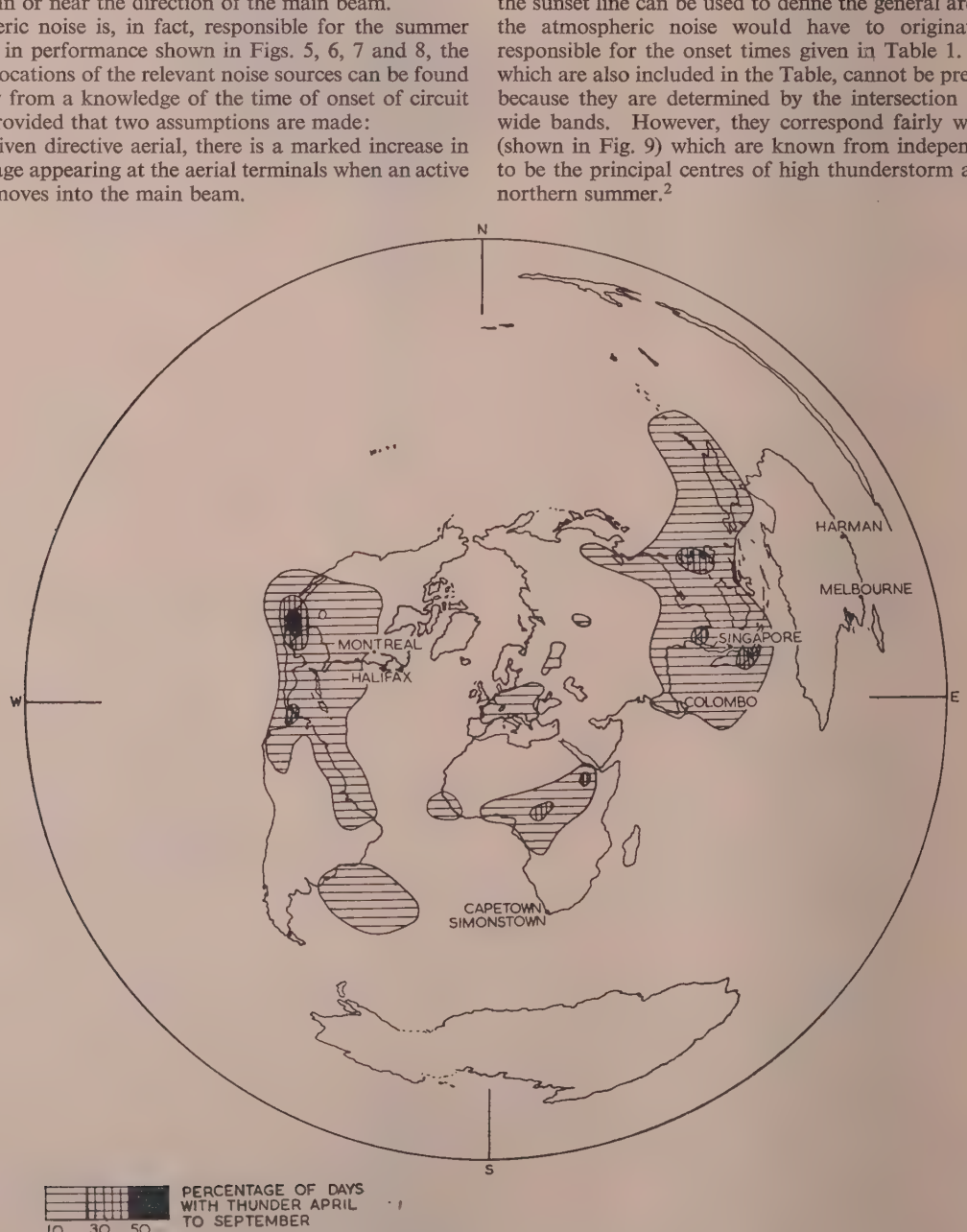


Fig. 9.—World distribution of thunderstorms, April–September (after Brooks).

f this argument is correct, it might be expected that there ild be a seasonal reversal of the asymmetry effects and that northern winter there would be an analogous fall in per- nance for reception at the terminals distant from London. wever, it is incorrect to think of winter and summer as causing omplete reversal of the characteristics of these circuits; it is portant to remember, e.g., that during northern winter the des between the sunset line and the great-circle routes are small ereas in northern summer they are much greater. Further- re, the latitude of London is greater than that of any of the thern hemisphere terminals. The net result is that circum- nces in northern summer are favourable to the control of the formance of these circuits by atmospheric noise; in northern ter the effect is masked by other factors.

A possible objection to any hypothesis that atmospherics are main cause of asymmetry is that the sources of atmospherics ght be expected to affect both the London and the distant ends the circuit in a similar manner. However, since no evidence the kind given in Fig. 10 is available for the distant receiving minals, it is not possible to resolve this difficulty at present.

(4) OBSERVED DATA ON ATMOSPHERIC INTERFERENCE

It is customary for operators to log the occasions when nospherics are detrimental to the reception of incoming ffic. When a number of circuits incoming from different rections terminate at the same receiving station, operational perience indicates that, in general, atmospherics originating in al storms affect all circuits simultaneously, but that local nospherics occur only rarely in comparison with those coming m distant sources. Furthermore, the dissimilarity of the rnal distribution of atmospheric interference for the two cuits referred to in Fig. 10 indicates that the great majority the atmospherics must have come from distant sources, the cation of which depends on the orientation of the receiving rial beams. In this connection Fig. 10(*a*) shows the frequency th which atmospherics were logged in London during reception signals from Capetown. The rapid rise in the incidence of ch log entries in summer after 1700 hours G.M.T. tends to pport the explanation already offered for the onset of asym- try as shown for summer in Fig. 5. In winter, when asymmetry this circuit is small, the detrimental effect of atmospherics is o small [see Fig. 10(*a*)]. For the Harman–London circuit, e onset of asymmetry occurs at about 1000 hours G.M.T. in mmer (see Fig. 8), and this is in fair agreement with the rise ring the forenoon in the number of atmospherics logged on the elbourne–London circuit in summer [see Fig. 10(*b*)].

The use of a directive receiving aerial can be shown to have important effect on the apparent diurnal changes in atmo- heric noise level. In Fig. 11 the difference in level of peak nospheric-noise voltages is shown for a rhombic and a half- ve vertical aerial. There is a pronounced increase in the ative level of the atmospherics received on the rhombic aerial er about 1600 hours G.M.T. and this is in accord with the idence given in Figs. 5 and 10(*a*).

Measurements of the diurnal change in the direction of rival of atmospherics in London have recently been made by ed,[4] who used a rotating cardioid aerial system. These show arly that, in the course of a day, the maximum number of nospherics is received from a direction which changes from st at 1200, through south at 2000, to west at 0300 hours G.M.T. ie obvious conclusion to be drawn from these measurements that the centre of gravity of the effective source of atmospherics oves from east to west with the sun, as implied by the data in ble 1.

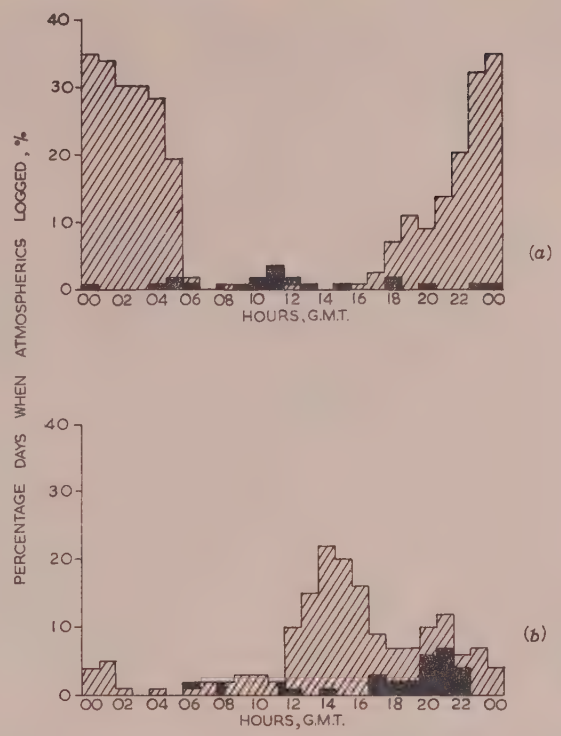


Fig. 10.—Percentage of occasions when atmospherics interfered with circuit operation, 1940–42.

(*a*) Reception from Capetown in London.
Summer, 1941.
Winter, 1941–42.
(*b*) Reception from Melbourne in London.
Summer, 1941.
Winter, 1940–41.

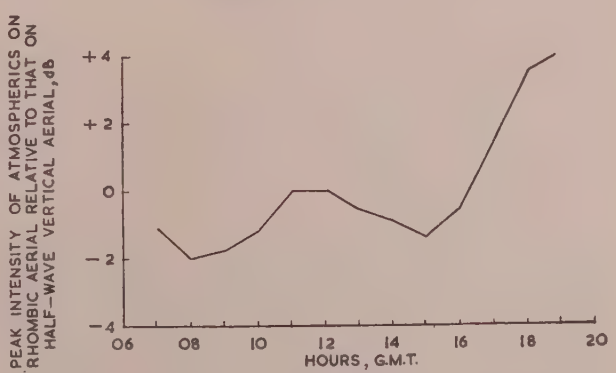


Fig. 11.—Peak atmospheric intensity on rhombic aerial relative to that on half-wave vertical aerial.

Rhombic aerial at Somerton (England) directed on Ascension Island; frequency 20·0 Mc/s; receiver bandwidth 2·5 kc/s; date 14.6.39.

(5) THE EFFECT OF ATMOSPHERICS ON THE CALCULATION OF LOWEST USEFUL HIGH FREQUENCY

The calculation of the lowest useful high frequency (l.u.f.) requires a knowledge of the noise voltage appearing at the terminals of the receiving aerial. If the conclusions reached in Section 4 are correct, the actual noise voltage, when a directive receiving aerial is used, will be determined mainly by the sources of atmospherics, often remote from the receiving site, lying in the direction of the main beam. Unfortunately, the noise grade charts normally

used for calculating l.u.f.'s are necessarily based on measurements made using omni-aerials. In view of this, for circuits which use directive aerials, it seems possible that some of the discrepancies between calculated values of l.u.f. and those observed in practice may be due to the use of noise levels which in these circumstances are inappropriate.

(6) CONCLUSIONS

On several trans-equatorial circuits connecting London with terminals to the south and east, it has been noticed that at certain times of day in northern summer, there is a fall in the relative performance of the circuit in the direction which involves reception in London. This has been attributed to the coincidence of the direction of the main beam of the receiving aerial with that of an equatorial area in which atmospherics may be generated, at the time of day when the atmospheric sources in this area are most active. In support of this conclusion it has been found that, on specific point-to-point circuits, the number of atmospherics logged by the operator does in fact begin to increase at about the times expected. It has also been shown that the difference between the atmospheric noise level for a directive receiving aerial and that for a vertical half-wave aerial increases significantly at a time of day corresponding to the passage of a distant source of atmospherics across the aerial beam. Thus, even if atmospheric noise does not account for all the asymmetry effects which have been reported from time to time, it is difficult to escape the conclusion that it is responsible for them in the cases considered here. However, it would seem that only by carrying out a series of careful simultaneous measurements at both ends of a number of circuits over a period of several years could a final conclusion be reached concerning the reasons for circuit asymmetry.

(7) ACKNOWLEDGMENTS

The authors wish to thank the Canadian Overseas Telecommunications Corporation and Cable and Wireless Ltd., for permission to use information relating to their circuits.

The paper is published by permission of the Admiralty and the Director of Radio Research of the Department of Scientific and Industrial Research.

(8) REFERENCES

(1) HUMBY, A. M., MINNIS, C. M., and HITCHCOCK, R. J.: 'Performance Characteristics of High-Frequency Radiotelegraph Circuits', *Proceedings I.E.E.*, Paper No. 1787 R, January, 1955 (**102** B, p. 513).

(2) BROOKS, C. E. P.: 'The Distribution of Thunderstorms over the Globe', Meteorological Office, London, Geophysical Memoir, No. 24, 1925.

(3) BUREAU, R., and BOST, R.: 'On Sources of Atmospherics', Notes préliminaires du Laboratoire National de Radioélectricité, Nos. 152 and 154.

(4) ISTED, G. A.: 'Irregularities in the E Region caused by Atmospheric Electricity', Report of Conference on the Physics of the Ionosphere at Cambridge, September, 1954, p. 150.

DISCUSSION ON

'HIGH-SPEED ELECTRONIC-ANALOGUE COMPUTING TECHNIQUES'*

Before the MERSEY AND NORTH WALES CENTRE *at* LIVERPOOL, 17*th October, and the* SOUTH-EAST SCOTLAND SUB-CENTRE *at* EDINBURGH, 20*th December*, 1955.

Mr. A. S. Aldred (*at Liverpool*): Far from being eclipsed by the digital machine, the analogue computer has a rightful place in research and industrial organizations as a complementary research tool. The advantages, of course, are that the quantities associated with the system being simulated, with which the engineer is familiar, are retained in the computer. The setting-up of a problem is a relatively simple procedure, i.e. the programming of the computer does not require the services of trained mathematicians.

Some doubt exists in my mind concerning the use of a basic amplifier consisting of a single stage of amplification terminated by a cathode-follower (Fig. 1). Tests carried out at Liverpool University on the solution of the s.h.m. equation indicated that with an amplifier of gain 1 000 considerable damping was introduced, but with amplifiers of gain 10^5 little or no damping was observed. Will the author's amplifier introduce damping in this way?

The methods outlined for discharging the integrator capacitors, resetting the integrators and switching the parameters are most ingenious. If I interpret Fig. 3 correctly, it refers to resetting integrators with the initial condition of zero output voltage only. What modification is necessary to reset initial conditions other than zero, or are the initial conditions set in at other points in the analogue-computer circuit?

I have recently encountered the problem of switching simultaneously the values of several parameters in an analogue computer, and have overcome it by introducing a brief 'hold' period. I envisage that the diode switches similar to those devised by the author could be used for this purpose.

The inclusion of function generators and multipliers in analogue computing circuits very often introduces errors, and attempts to improve accuracy and develop new techniques would be fruitful research projects. The advantages gained by including the diode networks in the feedback circuits of the amplifiers in Fig. 6 are partially offset by the introduction of bias voltages which prevents the inputs to the amplifiers going negative. Thus for small values of variables x and y the product term will be small compared with the $4c_1c_2y$ term; since this is removed by subtraction, large errors could occur.

* MACKAY, D. M.: Paper No. 1738 M, October, 1954 (see **102** B, p. 609).

Will the author indicate the application of the 4-dimensional lay to practical problems?

Mr. Langman (*at Liverpool*): At first sight this type of nputer has many advantages over the real-time computer. re are, however, limitations due to difficulties of program- g, coupled with the amount of useful information that can handled. The electrical industry is particularly interested in ng analogue computers to solve problems associated with sed-loop feedback systems. Operators of these computers y approach the problem by two methods: the first is to simulate system, component by component, and the second to use mathematical approach of representing directly the system's erential equations.

One of the main uses of an analogue computer is to optimize the ticular system under analysis, and the use of the machine cribed is obvious in this respect. Has the author thought of ng this type of repetitive computer on such problems, and, if what does he think about the possibility of programming the nputer for optimizing several parameters?

For control-system problems the time delay associated with amplifier will have to be taken into account at the design ge of the analogue. How much more difficult would be the ign of an analogue with such a computer than with the sting type of real-time computer?

Dr. J. P. Corbett (*at Liverpool*): The way in which one displays transients resulting from an analogue computer on a cathode-tube depends to a considerable extent upon the criterion to applied for optimum performance of the control system or ner dynamic system under examination. There are two possible thods of attack: displaying a number of traces at one time, d displaying a single transient which is continually changing shape. At a recent discussion on servo-system error criteria* o criteria emerged as being suitable: the minimization of the n.s. error in the servo system and the minimization of the egral of the product of error squared and time. In addition displaying a large number of transients and selecting from se the one most satisfactory, one could go through the possible mbinations of parameters and, by suitable switching, display ly a single transient, but by means of additional feedback in analogue itself cause the system to self-optimize automatically. Finally, will the author enlarge on his statement that obably not more than 10% of the available information pacity is utilized by the apparatus in its present form'?

Mr. R. A. Sheppard (*at Edinburgh*): In an automatic-control plication using an analogue computer at limiting accuracy, I mined the possibility of substituting a digital computer for the alogue computer. The former performs its calculations in crete steps, and it is desirable to know the time interval ween successive calculations which can be allowed without rming the quality of control. In my case, small size was more portant than calculating time. I set up the control-loop uations on a general-purpose digital machine and carried out trial runs with different values of the parameters. I believe that I might have obtained an approximate solution very much more quickly by the application of the author's technique, had suitable equipment been available. It is not always practicable to build special test-gear for single problems, and while we are grateful to the author for describing an improvement in technique, we should not altogether dismiss the complementary problem—that of helping what we might call the 'occasional user' to reap some benefit from the author's work.

Dr. D. M. MacKay (*in reply*): The inaccuracy due to damping, mentioned by Mr. Aldred, will depend on the number of cycles of the solution between each resetting period, and we may have escaped trouble because this number is small in most of our applications. It should be remembered, however, that the feed-back control and/or the neutralizing trimmer can be adjusted by inspection so as to eliminate damping from this and other causes at low and high frequencies respectively, if the required integrators, etc., are connected temporarily to solve the s.h.m. equation.

We normally set our initial conditions by supplying a steady current to each integrator through an additional input resistor, rather than by adjusting the initial level of the output of the preceding stage.

The subtraction of $4c_1c_2y$ in the multiplier requires careful setting of R_{13} (Fig. 6); but, in practice, it is the absolute error in the output, rather than the percentage error, which usually matters, since the scale is always adjusted so that the multiplier output swings over its full range, and errors when xy is small make no greater contribution to an integral than when it is large.

The most likely practical application of a 4-dimensional display would be to problems of the sort mentioned by Mr. Langman, where many parameters must be optimized by systematic trial and error. Brightness control can in such cases give a useful means of indicating the more satisfactory combinations and eliminating superfluous detail from the potentially confusing picture.

True time delays without frequency distortion are, of course, difficult to simulate electronically; they usually arise in problems of a different class from ours, more suitable for a real-time computer; but where a delay of only a few logons is required, standard multi-section delay networks could readily be used in a high-speed analogue. Where 'delay' means only sluggishness representable by a lagging transfer function, of course, no difficulty arises.

My estimate of informational efficiency was a rough one based on a comparison of the accuracy of solutions and their logon content with the information rate which the same electronic components could handle as an amplifier.

Many of the facilities desired by Mr. Sheppard could, in principle, be added to a digital computer if its programming arrangements were suitably modified; but quite a wide range of problems could be catered for by a 'universal' analogue computer along the lines I have described—if someone were willing to build it and make it available.

* Westcott, J. H.: 'The Minimum-Moment-of-Error Squared Criterion: a New ormance Criterion for Servo Mechanisms', *Proceedings I.E.E.*, Paper No. 1644 M, ch, 1954 (see 101, Part II, p. 471).

621.396.933.2:621.396.11

Paper No. 2124 [illegible]
July 19[illegible]

FLUCTUATIONS IN CONTINUOUS-WAVE RADIO BEARINGS AT HIGH FREQUENCIES

By W. C. BAIN, M.A., B.Sc., Ph.D.

(*The paper was first received 2nd February, and in revised form 7th April*, 1956.)

SUMMARY

Bearing observations have been made with an Adcock direction-finder on distant transmitters in the 3–4 Mc/s range, and the auto-correlation function of their time variation has been calculated. Curves of the form $\varepsilon^{-\tau/\tau_0}$ have been fitted to the functions obtained, the resulting values of τ_0 having a mean of 0·81 sec. The results differ from these in the 6–20 Mc/s band in that the standard deviation in a group of observations is not correlated significantly with the value of τ_0.

(1) INTRODUCTION

In a previous paper,* results have been reported on the fluctuations of radio bearing on an Adcock direction-finder in the frequency range 6–20 Mc/s. This work has now been extended to cover the low-frequency end of the h.f. band to determine whether any alteration takes place there in the nature of the fluctuations, with particular reference to their time scale. Measurements of this type can be used in estimating what improvement in bearing accuracy may be expected from averaging bearings taken over a short period, as is described in the earlier paper.

(2) RESULTS

On various occasions during October and November, 1954, photographic observations were made in rapid succession of the screen of a cathode-ray direction-finder in the manner described in the earlier paper. This work was carried out between sunset and sunrise to obtain reasonably low absorption losses on frequencies below 4 Mc/s. Details of the transmitters observed are given in Table 1.

Table 1

STATIONS OBSERVED

Location of transmitter	Frequency	Bearing from Slough	Distance
	Mc/s	deg	km
Frankfurt ..	3·188	99	670
Rome ..	3·995	132	1470

Nine groups of observations were taken, and were analyse as in the previous paper,* i.e. the autocorrelation functions c the bearing observations in each group were calculated. Curve of the form $\varepsilon^{-\tau/\tau_0}$ were fitted to them, τ being the time variabl and τ_0 being adjusted to give the best fit to each functio The values of τ_0 thus obtained were distributed as shown i Table 2.

Comparison with Fig. 2 of the previous paper shows that th distribution contains a rather larger proportion of high valu of τ_0 than is the case with the higher-frequency stations, for whic 90% of the values obtained lie below 1·2 sec. There is litt difference in the means for the two sets of transmitters, whic are 0·81 sec in the present experiments and 0·75 sec in th previous results. However, there is a marked difference for th standard deviation of an observation in each group, since the is now no significant negative correlation between τ_0 and th standard deviation. Indeed, the values of standard deviation f the groups with τ_0 greater than 1·0 sec are all above 5 deg. Th is very different from the previous results, none of which co tained slow fluctuations of large magnitude.

It must therefore be concluded that the results previous obtained in the range 6–20 Mc/s cannot be safely extended cover frequencies in the range 3–4 Mc/s. At these lower fr quencies, slower rates of bearing variation (corresponding to in the range 1–2 sec) appear to be dominant at times, the fluctu tions being often of considerable amplitude. The mean val of τ_0 derived from the data at 3–4 Mc/s is 0·81 sec.

It is, of course, possible that as the frequency moves fro 20 Mc/s down to 6 Mc/s there is a trend towards slower fluctu tions which is concealed by the small number of samples take However, to a first approximation, no allowance need be ma for this in calculations of time averages. The change behaviour below 6 Mc/s is apparent even from the few sampl and is therefore almost certainly more pronounced. It may associated with the fact that night conditions were being studi here, as opposed to day-time conditions in the higher-frequen experiments.

Table 2

NUMBER OF VALUES OF τ_0 FALLING WITHIN A GIVEN RANGE

Range of τ_0, sec ..	0–0·2	0·2–0·4	0·4–0·6	0·6–0·8	0·8–1·0	1·0–1·2	1·2–1·4	1·4–1·6	1·6–1·8	1·8–2·0
Number of values	1	2	2	0	0	1	1	1	1	0

(3) ACKNOWLEDGMENTS

The author acknowledges the assistance of Mr. J. Bell in t observations and computations. The work described was carri out as part of the programme of the Radio Research Boa This paper is published by permission of the Director Radio Research of the Department of Scientific and Industr Research.

* BAIN, W. C.: 'On the Rapidity of Fluctuations in Continuous-Wave Radio Bearings at High Frequencies', *Proceedings I.E.E.*, Paper 1715 R, October, 1954 (102 B, p. 541).

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.
The paper is an official communication from the Radio Research Station, Department of Scientific and Industrial Research.

PROCEEDINGS OF THE INSTITUTION OF ELECTRICAL ENGINEERS

Part B. RADIO AND ELECTRONIC ENGINEERING (INCLUDING COMMUNICATION ENGINEERING) JULY 1956

CONTENTS

	PAGE
The Calibration of Inductance Standards at Radio Frequencies.....L. HARTSHORN, D.Sc., and J. J. DENTON, B.Sc.	429
Nickel–Chromium–Aluminium–Copper Resistance Wire.....A. H. M. ARNOLD, Ph.D., D.Eng.	439
Discussion on 'An Extended Analysis of Echo Distortion in the F.M. Transmission of Frequency Division Multiplex'.....	447
An On-Off Servo Mechanism with Predicted Change-Over.....J. F. COALES, O.B.E., M.A., and A. R. M. NOTON, B.Sc.	449
The Dual-Input Describing Function and its Use in the Analysis of Non-Linear Feedback Systems J. C. WEST, Ph.D., J. L. DOUCE, Ph.D., and R. K. LIVESLEY, Ph.D.	463
Discussion on 'Artificial Reverberation'	474
The Effect upon Pulse Response of Delay Variation at Low and Middle Frequencies.....M. V. CALLENDAR, M.A.	475
An Electronic Machine for Statistical Particle Analysis.....H. N. COATES, Ph.D., B.Sc.(Eng.)	479
A Ferrite Microwave Modulator employing Feedback.....W. W. H. CLARKE, Ph.D., B.Sc., W. M. Searle, M.Sc., and F. T. Vail, B.Sc.	485
Wide-Band Noise Sources using Cylindrical Gas-Discharge Tubes in Two-Conductor Lines.....R. I. SKINNER, B.E.	491
The Application of Transistors to the Trigger, Ratemeter and Power-Supply Circuits of Radiation Monitors E. FRANKLIN, Ph.D., and J. B. JAMES	497
A Point-Contact Transistor Scaling Circuit with 0.4 microsec Resolution..... G. B. B. CHAPLIN, M.Sc., Ph.D.	505
A Junction-Transistor Scaling Circuit with 2 microsec Resolution .....G. B. B. CHAPLIN, M.Sc., Ph.D., and A. R. OWENS, M.Sc.	510
Discussion on the above three Papers	516
Frequency-Modulation Radar for Use in the Mercantile Marine.....D. N. KEEP, B.Sc.(Eng.)	519
Change of Phase with Distance of a Low-Frequency Ground Wave Propagated across a Coast-Line B. G. PRESSEY, M.Sc.(Eng.), Ph.D., G. E. ASHWELL, B.Sc., and C. S. FOWLER	527
The Deviation of Low-Frequency Ground Waves at a Coast-Line.....B. G. PRESSEY, M.Sc.(Eng.), Ph.D., and G. E. ASHWELL, B.Sc.	535
The Propagation of a Radio Atmospheric.....C. M. SRIVASTAVA, M.Sc.	542
The Prediction of Maximum Usable Frequencies for Radiocommunication over a Transequatorial Path.....G. McK. ALLCOCK, M.Sc.	547
Asymmetry in the Performance of High-Frequency Radiotelegraph Circuits.....A. M. HUMBY and C. M. MINNIS, M.Sc.	553
Discussion on 'High-Speed Electronic-Analogue Computing Techniques'.....	558
Fluctuations in Continuous-Wave Radio Bearings at High Frequencies.....W. C. BAIN, M.A., B.Sc., Ph.D.	560

Bibliographical References.—It is requested that bibliographical reference to an Institution paper should always include the serial number of the paper and the month and year of publication, which will be found at the top right-hand corner of the first page of the paper. This information should precede the reference to the Volume and Part.
Example.—SMITH, J.: "Reflections from the Ionosphere," *Proceedings I.E.E.*, Paper No. 3001 R, December, 1954 (102 B, p. 1234).

Published by The Institution, Savoy Place, London, W.C.2. *Telephone:* Temple Bar 7676. *Telegrams:* "Voltampere, Phone, London."
Printed by Unwin Brothers Limited, Woking and London.